TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Aplicació de mètodes de regressió robusta en els models de predicció

Autor: Aleix Salvador Barrera

Director: Francesc Carmona Pontaque

Departament: Genètica, Microbiologia i Estadística

Convocatòria: Juny 2019

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

Un dels problemes més usuals al món de l'estadística consisteix en explicar com es relaciona una variable d'interès Y amb una covariable o una sèrie de covariables explicatives X. En l'àmbit de la regressió, el mètode dels mínims quadrats ordinaris (MQO) és l'òptim quan es compleixen les suposicions necessàries. però si alguna o algunes d'aquestes suposicions no es compleixen, la regressió per mínims quadrats ordinaris pot funcionar malament i passa a ser un mètode invàlid. La regressió robusta és una alternativa als mínims quadrats ordinaris que utilitza suposicions menys restrictives i que estima els coeficients de la regressió de molt millor forma quan els valors atípics estan presents a les dades. Els valors atípics violen la suposició de la distribució normal dels residus i provoquen un fort biaix als coeficients estimats per mínims quadrats ordinaris ja que obtenen més influència de la que mereixen. Actualment existeixen diferents mètodes de regressió robusta, però aquest treball es centrarà en la regressió quantil lineal, en la regressió que es basa en la hipòtesi d'utilitzar la distribució t-Student en comptes de la Normal i finalment en la regressió per mínims quadrats ponderats.

PARAULES CLAU

Regressió, mínims quadrats ordinaris, suposicions, regressió robusta, valors atípics, biaix, influència, quantils, distribució t-Student.

ABSTRACT

One of the most common problems in statistics consists in explain how an interest variable Y is related to an explanatory variable or some explanatory variables X. In Regression, Ordinary Least Squares (OLS) is optimum when necessary assumptions are valid. However, when some of these assumptions are invalid, ordinary least squares regression can work wrong and becomes to an invalid method. Robust regression provides an alternative to ordinary least squares regression that works with less restrictive assumptions and provides much better regression coefficient estimates when outliers are present in the data. Outliers violate the assumption of normally distributed residuals in least squares regression and they tend to distort the least squares coefficients by having more influence than they deserve. Currently there are many different robust regression methods, but this memory will focus in quantile regression, robust statistical modeling using the t distribution instead of Normal distribution and finally weighted least squares.

KEY WORDS

Ordinary Least Squares, assumptions, robust regression, outliers, influence, quantile, t distribution.

CLASSIFICACIÓ AMS

- 62F35 Robustness and adaptive procedures
- 62J05 Linear regression
- 62-04 Explicit machine computation and programs (not the theory of computation or programming)
- 62-07 Data analysis

ÍNDEX

1. INTRODUCCIÓ

L'estadística és la ciència que permet prendre decisions davant de situacions d'incertesa. Aquestes decisions es basen en inferir a partir de mostres, tant per estimar valors poblacionals com per a realitzar proves que consisteixen en contrasts d'hipòtesis. En l'àmbit de l'estadística s'han desenvolupat dos tipus de proves: les proves paramètriques i les no paramètriques.

Per una banda, les proves paramètriques tenen en compte els paràmetres de la població i requereixen per a la seva utilització que es compleixin una sèrie d'assumpcions, ja que si no es compleixen poden portar a conclusions errònies i per tant es poden considerar com a proves no vàlides. Tanmateix, però, sota aquestes assumpcions s'obtenen procediments òptims. Per exemple, en el cas de la regressió el procediment òptim és el dels mínims quadrats, i per a models paramètrics en general, els procediments òptims clàssics són els estimadors basats en la màxima versemblança. De forma molt general, les condicions que cal que es compleixin per tal d'utilitzar les proves paramètriques són les següents: dades incorrelacionades (independents unes de les altres), variabilitat constant (homocedasticitat), distribució Normal de les variables poblacionals, linealitat i escala de mesura d'interval o raó.

Per altra banda, les proves no paramètriques no necessiten el compliment de cap d'aquestes condicions que s'acaben d'exposar, ja que es poden utilitzar per analitzar també variables nominals i ordinals, la distribució poblacional pot ser qualsevol i no és necessari suposar res respecte les variàncies poblacionals ni respecte la linealitat (això últim en la majoria dels casos).

Les proves paramètriques són més potents que les no paramètriques quan es compleixen les condicions necessàries, no obstant, quan aquestes condicions no s'assoleixen augmenta la probabilitat d'arribar a conclusions errònies i per tant, la potència d'aquestes proves es veu disminuïda notablement.

Donat que les condicions no sempre es compleixen, una possible estratègia per tal d'assegurar en la major part dels casos la màxima potència, seria la d'utilitzar sempre les proves no paramètriques ja que si es compleixen les condicions d'aplicació de les proves paramètriques, la pèrdua de potència no és molt gran (és conegut que els mètodes no paramètrics tenen una alta potència quan es compleixen les condicions d'aplicació dels paramètrics i que tenen molt poques probabilitats d'arribar a conclusions diferents de les obtingudes pels mètodes paramètrics tradicionals), i si no és compleixen, són aquestes les proves que cal dur a terme. Una altra alternativa és la utilització dels mètodes robustos, els quals són menys potents que els paramètrics però superen als no paramètrics clàssics. El major punt a favor d'aquests mètodes és que no es veuen afectats per l'existència de valors atípics o *outliers* i que no requereixen el compliment d'algunes de les condicions d'aplicació de les proves paramètriques. Per tant aquests mètodes poden ser d'utilitat per a la realització d'inferències i/o prediccions sense la necessitat de depurar les dades extremes ja que estan dissenyats per a realitzar prediccions sobre el model, reduïnt la possible influència que podria tindre la presència de valors anòmals.

1.1. JUSTIFICACIÓ

La meva motivació per fer aquest treball va vindre pel meu gran interès pels mètodes de predicció, en concret la regressió lineal. Per aquest motiu vaig contactar amb el professor de Models Lineals i li vaig explicar la situació. Ell em va proposar la regressió robusta, ja que és un tema que cada cop es fa servir més en els anàlisis de predicció. La veritat és que des del primer moment em va semblar un tema tant interessant com útil, ja que malauradament les dades del món real no són "perfectes" i la majoria dels cops no s'ajusten a les assumpcions necessàries per dur a terme els mètodes clàssics de l'estadística paramètrica.

1.2. OBJECTIUS

Els objectius d'aquest treball són:

- Definir el concepte de robustesa com a propietat dels estimadors estadístics i traslladar-lo a l'àmbit de la regressió.

- Resumir l'estat actual de la regressió robusta i exposar els casos en que és interessant dur-la a terme per aconseguir una major comprensió de la seva utilitat.

- Utilitzar tres mètodes de regressió robusta per analitzar una base de dades reals i comparar-ne els resultats extrets amb els de la regressió lineal simple.

1.3. METODOLOGIA

Aquest treball està dividit en dos parts clarament diferenciades.

La primera, purament teòrica, està enfocada en fer un resum global dels conceptes necessàris per seguir endavant amb el treball, com per exemple el concepte de robustesa o els fonaments de la regressió lineal. També es centra en detallar quins són els mètodes de regressió robusta que s'explicaran al treball.

La segona és la part pràctica del treball i és on es posaran en pràctica els mètodes explicats anteriorment per tal d'analitzar una base de dades. Aquesta base de dades es troba a la web *Machine Learning Repository*, l'enllaç de la qual és el següent: http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

Destacar que el software informàtic que s'utilitzarà al llarg de tot el treball per dur a terme els mètodes de regressió robusta exposats i analitzar les dades és l'R. En particular, s'utilitzaran les funcions "rq", "treg" i "rlm" de les llibreries "quantreg", "SMIR" i "MASS" respectivament.

1.4. ESTRUCTURA

La memòria del treball està estructurada en 7 capítols.

El primer capítol conté la introducció del treball així com la seva justificació, els objectius, la metodologia emprada i els agraïments.

Al segon capítol s'introdueix l'estimació estadística, els tipus d'estimadors i les propietats d'aquests, entre les quals apareix el concepte de robustesa.

Seguidament, al tercer capítol s'engloben les tècniques d'anàlisi predictiu per tal de situar la regressió dins del context d'aquestes tècniques.

Al quart capítol, es detalla el funcionament de la regressió lineal i les mancances que té davant la violació de les suposicions necessàries per la seva utilització. A part, també s'introdueixen els mètodes de regressió robusta capaços de solucionar aquestes mancances.

Al cinqué capítol s'entra en detall en cadascun dels tres mètodes de regressió robusta inclosos al treball, explicant el seu funcionament i proporcionant algun exemple gràfic d'aquest.

Al sisé capítol es realitza la part més pràctica del treball, ja que s'aplica la regressió lineal i els mètodes de regressió robusta prèviament explicats a una base de dades real mitjançant R.

Finalment, al sèptim capítol s'exposen les conclusions extretes del treball.

1.5. AGRAÏMENTS

Vull agrair al meu tutor, Francesc Carmona, l'ajuda, el seguiment i l'orientació que m'ha donat durant tot el període en que he dut a terme el treball de fi de grau.

2. ESTIMACIÓ

2.1. INTRODUCCIÓ

L'estimació és un procés estadístic que permet establir conclusions sobre característiques poblacionals a partir dels resultats d'una o més mostres.

Donada una variable aleatòria X que defineix una població (discreta o contínua), es defineix com a θ una característica d'interès d'aquesta població que s'anomena paràmetre poblacional, la qual és constant i generalment desconeguda, i es desitjarà estimar-la. Es pot donar el cas en que no només hi hagi una característica d'interès, sinó que es vulgui conèixer, per exemple, la mitjana, la mediana i la variància d'una població en concret. En aquesta situació es tindrà un vector en R^3 que inclourà els tres paràmetres a estimar. Generalment es definirà un vector Θ en R^p on p serà el nombre de paràmetres a estimar.

Per tal d'estimar la característica o característiques d'interès de qualsevol població es faran servir estimadors, els quals es poden definir com a una funció de les dades mostrals, o dit d'una altra manera, com a una fòrmula que depèn dels valors obtinguts d'una mostra. Cal destacar que hi ha dos tipus d'estimadors: els estimadors puntuals i els estimadors per intervals.

Els estimadors puntuals són els que assignen al paràmetre directament el valor obtingut de l'estimació. En efecte, existeixen molts possibles estimadors d'un paràmetre poblacional, ja que, per exemple, per tal d'estimar la mitjana μ d'una població es pot utilitzar la mitjana mostral igual que la mediana o la moda. Per aquest fet és necessàri donar algunes propietats desitjables als estimadors amb l'objectiu d'ajudar a decidir quin és el més adequat en cada cas. Més endavant, al punt 2.2 (Propietats dels estimadors) es tractaran les característiques dels estimadors que ens permetran escollir entre un estimador puntual o un altre segons quines siguen les circumstàncies de l'estudi.

Altrament, els estimadors per intervals estan determinats per dos valors dins dels quals s'afirma que es troba el veritable valor del paràmetre poblacional amb certa probabilitat. Es pot dir que són uns límits que es donen al valor estimat per tal d'afirmar, sota un criteri de probabilitat, que el veritable valor no els sobrepassarà. Aquest tipus d'estimadors són del tipus $\theta_1 < \theta < \theta_2$, on θ és el paràmetre a estimar i es trobarà entre θ_1 i θ_2 amb un determinat nivell de confiança.

2.2. PROPIETATS DELS ESTIMADORS

Les propietats o característiques que determinen si es té un bon estimador són les següents:

- **Biaix**: Un estimador θ^* del paràmetre θ és no esbiaixat si la seva esperança és igual al paràmetre poblacional. És a dir:

$$E[\theta^*] = \theta$$

Alguns estimadors no compliran aquesta propietat, són els denominats estimadors esbiaixats. Direm que θ^* és un estimador esbiaixat del paràmetre θ si es compleix que:

$$E[\theta^*] = \theta + b(\theta)$$

on $b(\theta)$ és el biaix de l'estimador.

- **Eficiència**: Donats dos estimadors θ_1^* i θ_2^* per a un mateix paràmetre poblacional θ, serà el més eficient el que tingui menor variància per a qualsevol mida mostral. Dit d'una altra manera, θ_1^* és més eficient que θ_2^* si:

$$Var(\theta_1^*) < Var(\theta_2^*)$$

- **Suficiència**: Un estimador és suficient quan s'utilitza tota la informació de la mostra al seu càlcul.

- **Consistència**: Un estimador és consistent quan coincideix amb el veritable valor del paràmetre que es vol estimar quan la mida mostral tendeix a infinit. És a dir, $\theta^*(x_1, x_2, \dots, x_n)$ (això significa que l'estimador θ^* s'obté amb les dades de la mostra $x_1, x_2, \dots, x_n$) és un estimador consistent del paràmetre θ si es compleix que:

$$\lim_{n\to\infty} \theta^*(x_1, x_2, \dots, x_n) = \theta$$

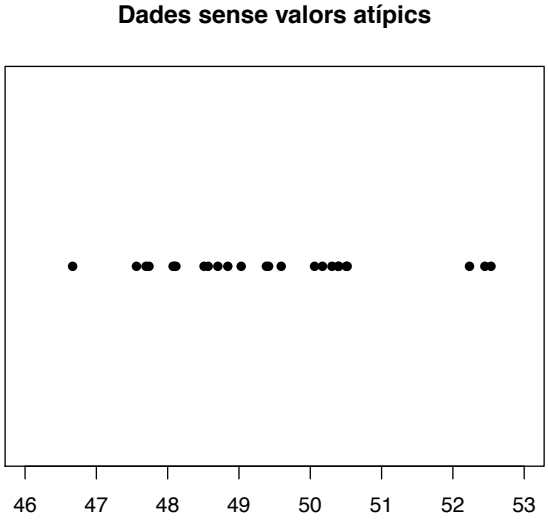
2.3. ROBUSTESA D'UN ESTIMADOR

D'acord amb el que s'ha explicat a l'apartat anterior, un paràmetre poblacional s'estima a partir de la informació extreta d'una mostra aleatòria d'aquesta mateixa població. Tanmateix, però, aquesta mostra pot estar contaminada per valors atípics (*outliers*) i/o per errors que afecten de forma negativa, ja que el seu comportament és molt diferent a la resta d'observacions. Un estimador es podrà qualificar com a robust quan no es veurà afectat per la presència d'aquests valors atípic (ni per la d'errors), és a dir, quan el seu valor sigui el mateix (o molt semblant) independentment de si la mostra conté, o no, valors atípics (o errors).

Per explicar de forma més clara i visual aquesta definició es proposa un exemple, en que es calcularà la mitjana i la mediana de dos mostres: la primera no contindrà cap valor atípic i la segona si.

D'entrada tenim una mostra de mida 25 d'una població que es distribueix normalment amb mitjana $\mu = 50$ i desviació típica $\sigma = 2$ (variància $\sigma^2 = 4$). La seva representació gràfica és la següent:

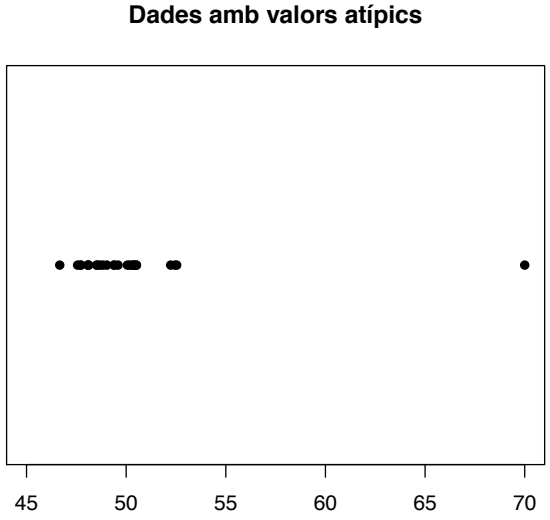
Imatge 2.1. Conjunt de dades generat de forma aleatòria sense cap valor atípic.



Es pot apreciar com clarament tots els valors es troben al voltant del 50. Utilitzant aquesta mostra la mitjana té un valor de 49.42 i la mediana de 49.38, tots dos valors bastant pròxims a la mitjana poblacional μ ($\mu = 50$), i per tant, tots dos estimadors són bons per estimar el veritable paràmetre poblacional en aquest cas.

Seguidament, afegim a la mostra anterior un valor atípic com per exemple el valor 70. La mostra quedaria representada així:

Imatge 2.2. Conjunt de dades generat de forma aleatòria amb un valor atípic.



En aquest cas es veu a simple vista que el valor que s'acaba d'introduir no es comporta de la mateixa manera que la resta de dades de la mostra ja que es troba molt allunyat dels altres. Tal com s'ha fet en el cas anterior, s'ha calculat la mitjana i la mediana per a aquesta nova mostra i els valors resultants han sigut 50.21 i 49.39 respectivament.

Taula 2.1.Comparació del valor de l'estimació de la mitjana i la mediana.

	Mitjana	Mediana
Mostra sense atípics	49.42	49.38
Mostra amb atípics	50.21	49.39
Diferència en valor absolut	0.79	0.01

A partir dels valors d'aquests dos estimadors amb cadascuna de les dos mostres es pot comprovar que el canvi que ha sofert la mitjana ha sigut molt superior al que ha sofert la mediana al introduir un valor atípic a la mostra. Per tant, aquest n'és un bon exemple de que la mediana és menys sensible a la presència de valors atípics que la mitjana aritmètica, la qual cosa mostra clarament la major robustesa que té (la mediana) com a estimador de la mitjana poblacional.

3. PREDICCIÓ

3.1. ANÀLISI PREDICTIU

L'anàlisi predictiu és una àrea de l'estadística que agrupa una gran varietat de tècniques de modelització, aprenentatge automàtic i mineria de dades, que analitza dades actuals i històriques per tal de fer prediccions del futur o d'esdeveniments desconeguts. Aquest conjunt de tècniques es basen en la identificació de relacions entre variables en situacions anteriors, amb l'objectiu d'aprofitar-les i predir possibles futurs resultats. No obstant això, cal tenir en consideració que la precisió dels resultats obtinguts depèn molt de la forma en que s'ha realitzat l'anàlisi de les dades, de la qualitat d'aquestes i de la qualitat de les suposicions en que s'ha basat l'estudi.

D'entrada pot semblar que l'anàlisi predictiu és el mateix que un pronòstic, però són dos coses totalment diferents. Així com un pronòstic realitza prediccions de forma més general, com per exemple, quants gelats seran venuts l'any que ve, l'anàlisi predictiu va més enllà i pot indicar quins individus és més probable que comprin un gelat. Aquesta informació té molt valor ja que si s'utilitza de la forma correcta, pot suposar un canvi radical en l'estudi o en el negoci perque permet orientar els esforços per ser més productius en la consecució dels objectius.

Per dur a terme l'anàlisi predictiu és necesita una gran quantitat de dades, tant de l'actualitat com del passat, per tal d'establir patrons de comportament i d'aquesta forma induir coneixement. Per exemple, en el cas anterior consistent en predir quins individus és més probable que es comprin un gelat, si es creuen dades referents a la temperatura registrada, l'època de l'any i si es cap de setmana es pot inferir quin perfil de persona menjarà gelats. Els ordinadors poden aprendre de forma autònoma (aprenentatge computacional) i així desenvolupar nous coneixements i capacitats. Per fer això només cal se'ls proporcioni el major recurs natural de la societat moderna: les dades.

3.2. TÈCNIQUES D'ANÀLISI PREDICTIU

Les tècniques que s'utilitzen per dur a terme l'anàlisi predictiu es poden dividir de forma molt general en dos grans grups: les tècniques de regressió i les tècniques d'aprenentatge computacional.

Per un costat, les tècniques de regressió són el pilar de l'anàlisi predictiu i es basen en la construcció d'una equació matemàtica que ajuda a representar les interaccions entre les diferents variables de l'estudi. Els mètodes més utilitzats dins d'aquest grup són la regressió lineal, l'anàlisi de supervivència i els arbres de decisió i regressió (anàlisi discriminant). Resumint breument cadascun d'aquests:

- **Regressió lineal**: és un dels mètodes mes utilitzats en l'estadística i analitza la relació existent entre la variable resposta (que normalment es nombra com a Y) i les variables predictores (que normalment es nombren com a X). Aquesta relació és expressada com a una funció lineal dels paràmetres, els quals s'ajusten per tal que l'error sigui mínim i la mesura d'ajust, òptima.

- **Anàlisi de supervivència:** analitza el temps que passa fins que succeeix un esdeveniment. Aquesta tècnica es va desenvolupar principalment en l'àmbit de la medicina però també s'utilitza en les ciències socials i en l'enginyeria, ja que els esdeveniments que s'analitzen poden ser tant la mort d'un pacient com la fallida d'una bombeta.

- **Arbres de classificació i/o regressió**: és un dels mètodes de l'anàlisi discriminant que es pot definir com a una alternativa al model de regressió, ja que intenta expressar una variable depenent o resposta com una combinació lineal de les variables independents o predictores. Com a diferència tenim que el model de regressió prediu

una variable numèrica mentre que l'anàlisi per arbres de classificació i/o regressió prediu una variable categòrica.

Per l'altre, l'aprenentatge computacional es va crear originalment per desenvolupar tècniques que permeteren als ordinadors aprendre i, com que avui en dia inclouen mètodes molt avançats de regressió i classificació, s'apliquen a diferents camps com per exemple diagnòstics mèdics, reconeixements facials i anàlisi del mercat de valors. Algunes de les tècniques més comuns dins d'aquest grup són:

- **Xarxes Neuronals:** són tècniques de modelat no lineal molt sofisticades que permeten modelar funcions complexes i s'utilitzen quan no es coneix la naturalesa exacta de la relació entre els valors d'entrada i els de sortida. Les xarxes neuronals aprenen la relació entre els valors d'entrada i sortida a través de l'entrenament, el qual pot ser per reforç, supervisat o no supervisat.

- **Màquines de vectors de suport (*Super Vector Machine* o *SVM*):** són tècniques utilitzades per detectar i explotar patrons complexes de les dades agrupant, ordenant i classificant-les. Es poden definir com a mètodes d'aprenentatge que s'utilitzen per realitzar classificacions binàries i estimacions de regressió.

- ***Naïve Bayes:*** el classificador bayesià ingenu està basat en la regla de la probabilitat condicional de Bayes (per això el seu nom) i assumeix que tots els predictors són independents, la qual cosa el converteix en una eina de classificació eficaç, que al mateix temps és fàcil d'interpretar.

- **K-Veïns més propers (*K-Nearest Neighbor* o kNN):** pertany als mètodes estadístics de reconeixement de patrons i no imposa a priori cap suposició sobre la distribució de la qual prové la mostra. Es fa servir un conjunt de dades d'entrenament amb valors positius i negatius per classificar noves observacions. Es consideren els "k" veïns més propers i s'utilitza el signe de la majoria d'aquests per classificar la nova mostra. El rendiment d'aquest mètode depèn, òbviament, de la mesura de distància utilitzada per localitzar els veïns més propers i del nombre de veïns utilitzats per classificar la nova mostra.

Arribats a aquest punt, la regressió lineal ha estat situada dins dels diferents mètodes d'anàlisi predictiu i també s'ha definit el concepte de robustesa com a propietat d'un estimador estadístic. A partir d'ara el treball es centrarà en juntar aquestes dos idees. Per una banda, s'exposarà el funcionament de la regressió lineal i les seves mancances davant la violació de

les condicions necessàries per la seva utilització. I per altra banda, es definirà la regressió robusta, com a una solució de les mancances que té la regressió lineal quan no es compleixen les condicions necessàries, efectuant una introducció, un resum de l'estat actual d'aquest tipus de regressió i exposant tres mètodes diferents per dur-la a terme.

4. REGRESSIÓ

4.1. REGRESSIÓ LINEAL

El model de regressió lineal consisteix, com s'ha explicat en anterioritat, en analitzar la relació que té una variable depenent o resposta (Y) amb una o més variables independents o predictores (X_1,X_2,...,X_k) desenvolupant una equació lineal amb finalitats predictives. El cas concret en que només hi ha una variable predictora s'anomena regressió lineal simple i quan n'hi ha més d'una, regressió lineal múltimple. Aquest apartat, però, es centrarà en el cas en que només es té una variable predictora per tal d'explicar el funcionament i les condicions necessàries per aplicar model de regressió lineal.

Generalment, la regressió lineal es sol formalitzar com la mitjana condicionada de la variable resposta en funció del valor que pren la variable explicativa o regressora, és a dir, $m(x) = E(y|X = x)$ per a cada possible valor *x* de la variable *X*. Per tant, la variable resposta es pot descomposar en funció del resultat de la variable regressora X més un terme d'error aleatori que es troba centrat al valor zero. La fòrmula és la següent:

$$y = m(x) + \epsilon$$

on ϵ és conegut com error i es comporta com una variable aleatòria no observable que conté la variabilitat no explicada per la variable regressora, és a dir, la que es deu als errors de mesura o a alguns altres factors que no es poden controlar.

A continuació, com que la regressió lineal és un mètode paràmetric, es presenten les condicions necessàries que ha de complir el terme de pertorbació ϵ per poder aplicar el model:

- Linealitat: la funció de regressió és una recta. Per tant, el model s'escriu com

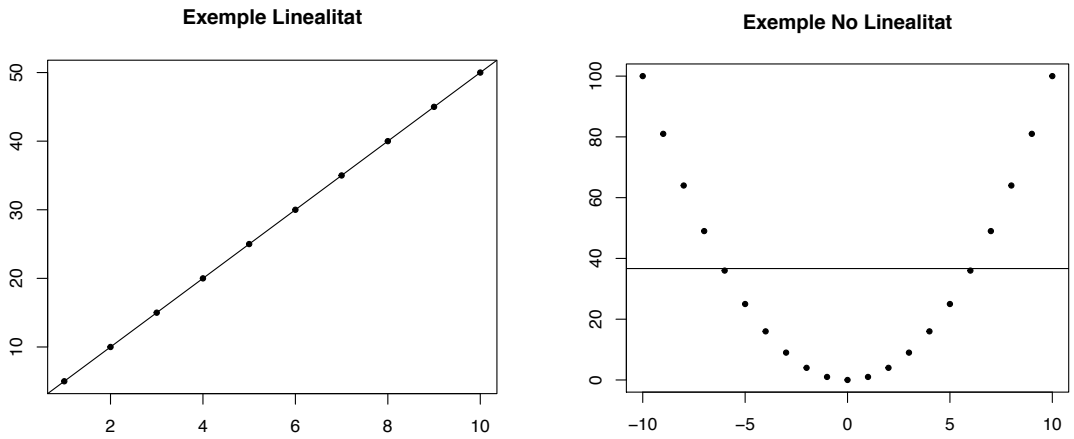
$$Y = m(x) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

  on β_0 i β_1 són paràmetres que caldrà estimar a partir d'una mostra. Per la qual cosa, cal que l'esperança del terme de pertorbació ϵ sigui igual a zero per a totes les observacions de l'estudi per tal d'evitar situacions on les dades segueixin patrons no lineals. És a dir:

$$E(\epsilon, X = x) = 0, \quad \forall x$$

A continuació s'adjuntaran dos gràfics mitjançant els quals es podrà observar un exemple de linealitat i un exemple en que no es compleix aquest supòsit:

Imatge 4.1. Exemple de linealitat i no linealitat en les dades.



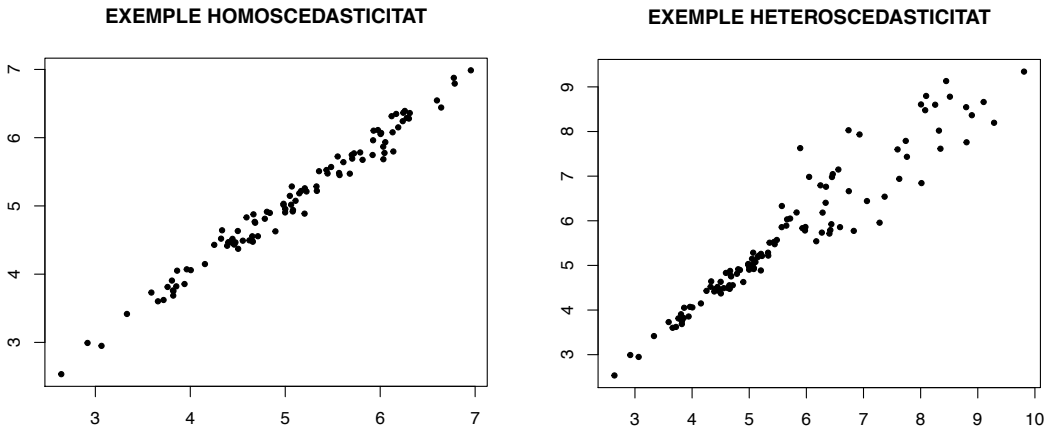
Cal destacar que la recta que uneix els punts en cadascun dels dos gràfics és la recta de regressió lineal calculada a partir del conjunt de dades utilitzat per realitzar cada gràfic.

- Homoscedasticitat: la variància de l'error es manté constant al llarg de la variable regressora, és a dir

$$Var(\epsilon|X = x) = \sigma^2, \quad \forall x.$$

Destacar que la situació en que no es compleix aquesta condició, i per tant, la variància no és constant, s'anomena heteroscedasticitat. Tot seguit s'adjunten gràficament les situacions d'homoscedasticitat i heteroscedasticitat:

Imatge 4.2. Exemple d'homoscedasticitat i heteroscedasticitat en les dades.



- Normalitat: el terme de pertorbació es distribueix normalment i es troba centrat al valor zero

$$\epsilon \sim N(0, \sigma^2), \quad \forall x.$$

- Observacions incorrelacionades: cal que les dades de la mostra siguin independents.

$$E(\epsilon_i \cdot \epsilon_j) = 0, \quad \forall i \neq j$$

Mitjançant la recta de regressió obtinguda a partir d'una mostra que compleix les condicions descrites en aquest apartat, es pot predir el valor de la variable resposta Y a partir d'un valor x de la variable regressora X, de forma que

$$\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_i, \quad \forall i \in \{1, \dots, n\}$$

A partir de les prediccions extretes s'obtenen els residus de la regressió, els quals es calculen de la següent forma:

$$\hat{\epsilon}_i = Y_i - (\widehat{\beta_0} + \widehat{\beta_1} X_i), \quad \forall i \in \{1, \dots, n\}$$

Per acabar amb aquest apartat, cal destacar que l'estimació dels paràmetres β_0 i β_1 es realitza mitjançant el mètode dels mínims quadrats ordinaris, el qual selecciona els paràmetres β_0 i β_1

tals que minimitzen la suma dels residus al quadrat. Per il.lustrar-ho millor, la funció a minimitzar és

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

així doncs, per minimitzar-la cal derivar i després igualar a zero. El resultat d'això són els estimadors dels mínims quadrats:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$
$$\widehat{\beta_1} = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

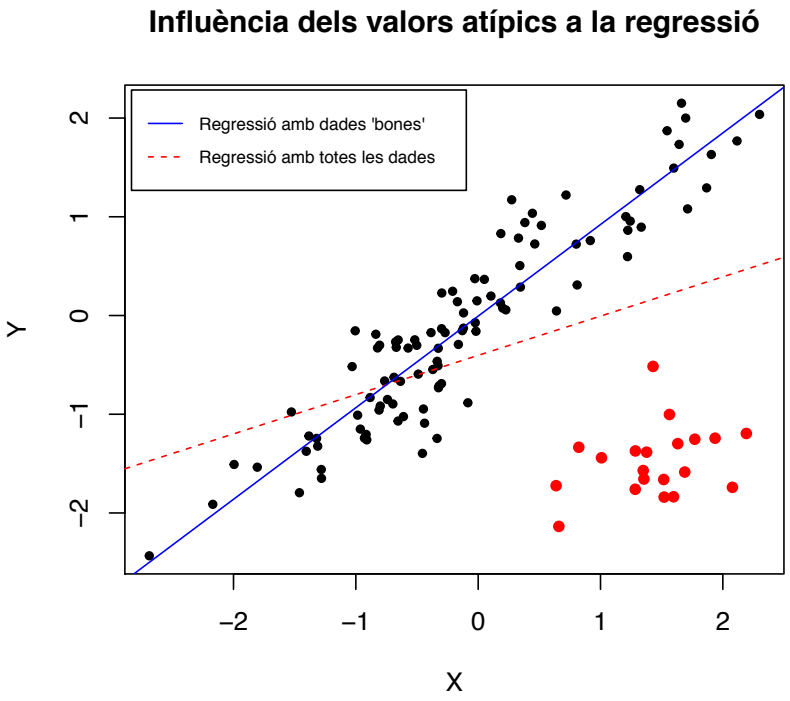
4.2. APLICACIÓ DE LA REGRESSIÓ ROBUSTA

En l'apartat anterior s'ha introduït la regressió lineal i ha sigut classificada dins dels mètodes paramètrics, ja que per la seva correcta utilització cal que el seu terme de pertorbació compleixi una sèrie de supòsits, entre els quals es troben la linealitat, la normalitat, l'homoscedasticitat... La violació de qualsevol d'aquests supòsits implica la realització d'un mètode estadístic que arriba a conclusions invàlides i que per tant no té sentit realitzar, ja que un no es pot fiar dels seus resultats perque es pot arribar a conclusions errònies.

Certament, un dels majors punts dèbils de la regressió lineal és la presència de valors atípics a la mostra d'estudi, els quals violen la hipòtesi de normalitat del terme de pertorbació, ja que aquests provoquen una forta desviació en les estimacions dels paràmetres. Per tant, els mètodes de regressió robusta s'apliquen per a donar una solució a aquesta alta vulnerabilitat que té la regressió lineal en front dels valors atípics.

Tot seguit, s'adjuntarà el gràfic d'un conjunt de dades (una variable resposta i una explicativa) generat amb l'R que conté 100 valors "bons" i 20 valors atípics que servirà per mostrar visualment la gran influència que tenen aquests sobre la regressió lineal, ja que per una banda es calcularà la regressió lineal sobre només el conjunt de les 100 dades "bones" i per altra, la regressió lineal sobre totes les dades incloent els atípics:

Imatge 4.3. Exemple de la influència dels valors atípics en la regressió lineal.



La gran desviació que es pot apreciar a simple vista de la recta de regressió vermella (regressió sobre tot el conjunt de dades) respecte de la blava (regressió sobre només les dades bones) és un gran exemple de la forta influència que tenen les dades atípiques sobre les estimacions dels paràmetres de la regressió lineal. Per tant, l'objectiu de la regressió robusta serà trobar l'estimació dels paràmetres que proporcionin una recta de regressió que s'assembli el màxim possible a la que no té en compte els valors atípics (la recta blava del gràfic anterior) però utilitzant totes les dades de la mostra.

4.3. ACTUALITAT DE LA REGRESSIÓ ROBUSTA

Generalment, els mètodes paràmetrics són els més utilitzats en els estudis estadístics, els quals, tal i com s'ha explicat en apartats anteriors, necessiten que es compleixin una sèrie de suposicions per tal de funcionar de forma correcta. Aquestes suposicions necessàries van provocar el desenvolupament de mètodes alternatius als paramètrics per tal d'obtenir uns resultats iguals o el més semblants possibles als d'aquests, però que treballin amb unes suposicions menys restrictives com en el cas de la regressió robusta, o directament sense cap suposició, com en el cas dels mètodes no paramètrics. Per tant, els mètodes de regressió robusta es podrien considerar com a mètodes semiparamètrics ja que treballen amb suposicions semblants a la regressió lineal (en quant a la linealitat, homoscedasticitat...), però n'hi ha una que no cal que es compleixi de forma tant estricta com a la regressió lineal. Aquesta

és la de la normalitat dels erros, i es tradueix com que els mètodes robustos treballen igual de bé tant si hi ha valors atípics com si no n'hi han.

Encara que des de que Huber va donar inici a la teoria quantitativa de la robustesa (la qual pretén quantificar l'estabilitat d'un estimador davant la presència de valors atípics a la mostra) l'any 1964 han passat més de 50 anys, aquesta encara no està completada ja que les corbes de biaix màxim (les quals mesuren la major discrepància possible entre el valor al qual convergeix un estimador i el valor del veritable paràmetre que es desitja estimar quan apareixen valors atípics a la mostra) dels estimadors de regressió robusta encara no es coneixen en la majoria dels casos. És per això que en l'actualitat es continua utilitzant el mètode dels mínims quadrats ordinaris quan es vol dur a terme un estudi per regressió, encara que les mancances d'aquest mètode davant l'incompliment de les seves suposicions siguin àmpliament conegudes. Tanmateix, però, els mètodes de regressió robusta s'estan fent servir cada cop més en el dia a dia dels anàlisis de predicció, ja que, tal i com es comprovarà més endavant en aquest treball, les seves prediccions difereixen molt poc respecte les que s'extreuen de la regressió lineal per mínims quadrats ordinaris sense tenir en compte les dades atípiques.

4.4. INTRODUCCIÓ DELS DIFERENTS MÈTODES DE REGRESSIÓ ROBUSTA A ANALITZAR

El següent punt a considerar, després d'introduir la regressió robusta i de fer un resum del seu estat actual, és el que presentarà els diferents mètodes a analitzar en aquest treball per tal d'aplicar aquest tipus de regressió. Aquests mètodes són tres:

- La regressió per quantils: aquest tipus de regressió consisteix en estimar els diferents quantils (com per exemple la mediana) d'una població i es diferencia de la regressió lineal per mínims quadrats ordinaris en que aquesta última es centra en estimar la mitjana.

- La regressió robusta t: aquest mètode es recolza sobre la hipòtesi d'utilitzar la distribució t-Student (en contes de la Normal) com a distribució del terme de pertorbació, com a conseqüència de la vulnerabilitat als valors atípics que tenen els mètodes estadístics basats en la llei Normal.

- La regressió per mínims quadrats ponderats: aquest mètode, a diferència de la regressió lineal per mínims quadrats ordinaris, assigna pesos diferents a les observacions de la mostra per tal de reduir la influència que tenen les dades atípiques en l'estimació.

5. MÈTODES DE REGRESSIÓ ROBUSTA

Aquest apartat es centrarà en desenvolupar detalladament cadascun dels mètodes de regressió robusta que han sigut presentats en l'apartat anterior. Es començarà per la regressió per quantils, seguida per la regressió amb la distribució t i es finalitzarà amb la regressió per mínims quadrats ponderats.

5.1. REGRESSIÓ QUANTIL LINEAL

El primer mètode a analitzar és la regressió quantil lineal. Aquest mètode consisteix en predir, tal i com el seu propi nom indica, un quantil determinat de la variable resposta i es diferencia de la regressió lineal en que l'objectiu d'aquesta és predir la mitjana de la variable Y.

Per definir de forma teòrica el concepte de quantil, es té que donada una variable aleatòria Y i un $\tau \in (0,1)$, el τ-èssim quantil és:

$$Q(\tau) = \inf\{Y: F(Y) \geq \tau\}$$

on F és la funció de distribució de Y.

Si per contra, es té una mostra amb observacions independents, també es pot trobar una estimació de la funció de distribució mitjançant la distribució empírica de la mostra. Aquesta es defineix com el quocient entre el nombre d'observacions inferiors o iguals al valor d'interès i el nombre total d'observacions:

$$\hat{F}(Y) = \frac{\#(Y_i \leq Y)}{n}$$

Un cop calculada la distribució empírica, es pot definir una estimació per als quantils de la mateixa forma que abans:

$$\hat{Q}(\tau) = \inf\{Y: \hat{F}(Y) \geq \tau\}$$

En altres paraules, per tal d'entendre-ho millor, el τ-èssim (cal recordar que $0 < \tau < 1$) quantil d'una distribució, és el punt que deixa una proporció τ de valors de la població per sota d'ell. Per exemple, el quantil d'ordre 0.50 (que és la mediana de la distribució) deixa un 50% de les observacions per sota d'ell, i el d'ordre 0.75, un 75%.

Efectivament es pot portar el concepte de quantil a la recta de regressió. Aquesta tècnica s'anomena regressió quantil lineal i té com a objectiu, a diferència de la regressió lineal, predir el quantil τ de la variable resposta Y. Així doncs, s'obté la següent equació:

$$Q_\tau(Y_i|X) = \beta_{0,\tau} + \beta_{1,\tau} \cdot X_i + \epsilon_{i,\tau}\,, \qquad \forall i \in \{1, \dots, n\}$$

on $\tau \in (0,1)$ i el τ-èssim quantil del terme de pertorbació és igual a zero ($Q_\tau(\epsilon_i|X) = 0$).

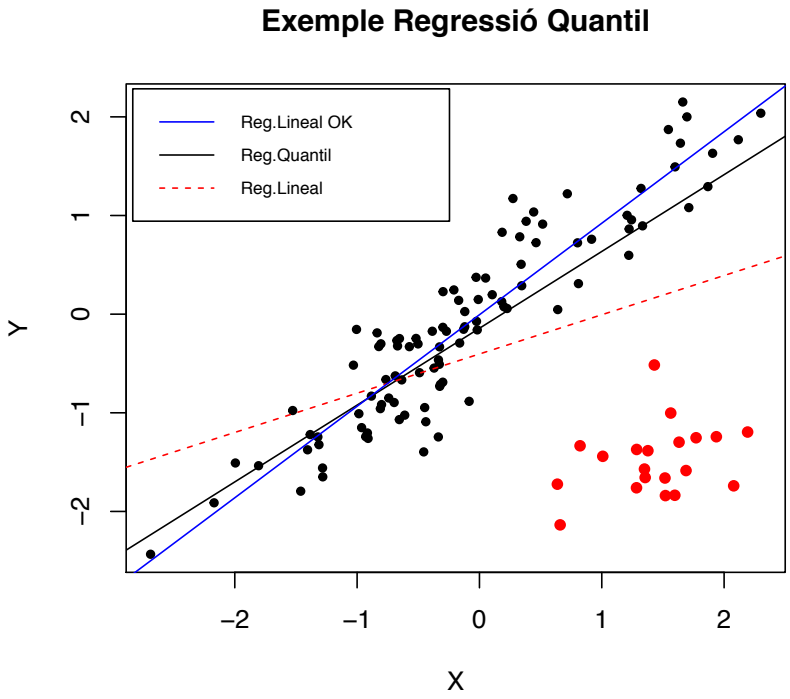
Les estimacions dels paràmetres $\beta_{0,\tau}$ i $\beta_{1,\tau}$ es troben mitjançant la minimització de la següent funció:

$$S = \left\{\sum_{Y_i \geq A} \tau \cdot \left|Y_i - \beta_{0,\tau} - \beta_{1,\tau} \cdot X_i\right| + \sum_{Y_i < A} (1-\tau) \cdot \left|Y_i - \beta_{0,\tau} - \beta_{1,\tau} \cdot X_i\right|\right\}$$

on $A = \beta_{0,\tau} + \beta_{1,\tau} \cdot X_i$.

Convé destacar que per tal d'obtenir les estimacions dels paràmetres es planteja el problema de minimització com un problema de programació lineal. Aquest aclariment és útil per comparar aquest mètode de regressió robusta amb el mètode de la regressió lineal en quan a l'estimació dels paràmetres, ja que per una banda, es pot apreciar que aquesta estimació consisteix en la minimització d'una funció objectiu en ambdós mètodes, però que per altra banda, en cadascun dels mètodes s'utilitza una tècnica diferent per dur-la a terme (els mínims quadrats en la regressió lineal i la programació lineal en la regressió quantil linea)l.
Tot seguit, per mostrar el funcionament i la potència d'aquest tipus de regressió, es mostrarà un exemple amb el conjunt de dades generat prèviament en l'apartat 4.2 (Aplicació de la regressió robusta) on hi apareixerà la recta de regressió quantil lineal, la recta de regressió lineal tenint en compte només les dades "bones" i la recta de regressió tenint en compte tot el conjunt de dades. Cal destacar que per dur a terme la regressió quantil lineal amb l'R s'utilitzarà la funció "rq" de la llibreria "quantreg" i també que dins d'aquesta funció hi ha un paràmetre que s'anomena "tau" que serveix per passar-li a la funció quin quantil de la variable resposta es vol predir. En aquest cas s'utilitzarà el valor de "tau" per defecte que és el quantil 0.5 (la mediana). El gràfic esmenat és el següent:

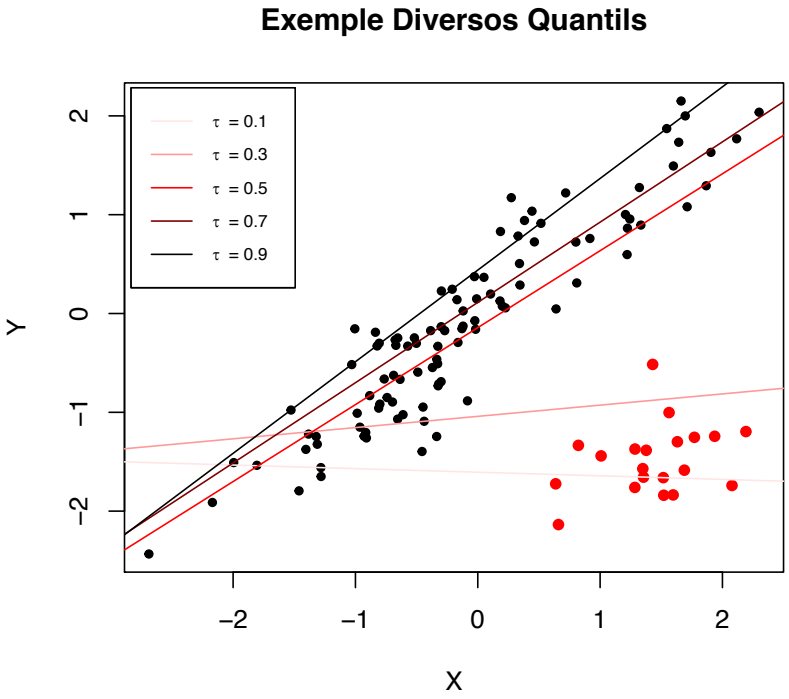
Imatge 5.1. Exemple d'aplicació de la regressió quantil lineal.



A partir del gràfic es pot apreciar clarament la forta influència que tenen els valors atípics sobre la recta de regressió lineal, ja que la diferència entre la recta blava (regressió lineal aplicada només als punts negres del gràfic) i la vermella (regressió lineal aplicada a tots els punts del gràfic) és més que evident. Per altra banda, la recta de regressió quantil lineal (recta negra) es veu molt poc afectada per la presència de valors atípics a la mostra ja que la diferència respecte la recta de regressió lineal aplicada només sobre les dades "bones" és mínima. Així doncs, gràcies a aquest exemple s'ha tornat a comprovar la major robustesa que té la mediana respecte la mitjana aritmètica, ja que com s'ha explicat abans, la regressió quantil lineal s'ha dut a terme sobre el quantil 0.50 (o sigui, la mediana) mentre que la regressió lineal tracta predir la mitjana.

Finalment, com que la regressió quantil lineal pot predir, no només el quantil 50 (mediana) sinó qualsevol dels quantils de la variable resposta, s'utilitzarà el conjunt de dades que s'ha generat per dur a terme l'exemple anterior i s'adjuntarà la representació gràfica de les rectes de regressió mesurant per a diferents quantils de la variable resposta Y:

Imatge 5.2. Exemple regressió quantil lineal predint diferents quantils.



Els quantils que s'han predit amb aquestes rectes són el 0.1, 0.3, 0.5, 0.7 i 0.9 i les rectes de regressió que els representen són les que es veuen al grafic, sent la que té un color vermell més clar (la que es troba més a baix) la que pertany al quantil 0.1 i la que té un color més fosc (la que es troba més a dalt) la que pertany al 0.9. Amb aquest gràfic es pot apreciar que els quantils més extrems, o sigui els més propers al 0 i a l'1 prediuen, com és d'esperar, valors allunyats del centre de la variable resposta, mentre que els quantils més propers al 0.5 o mediana, prediuen valors centrals de la variable resposta i són els que s'apropen més a les prediccions que es podrien extreure amb la regressió lineal si no hi haguessin valors atípics.

5.2. REGRESSIÓ ROBUSTA T

El segon mètode a analitzar és el de la regressió robusta utilitzant la distribució t-Student o la regressió robusta t. Aquest mètode, igual que l'anterior, consisteix en trobar la forma de restar-li importància a les observacions extremes o atípiques per tal que aquestes no influeixin tant en la recta de regressió i així, aconseguir que les prediccions siguin fiables malgrat la seva presència.

Convé destacar que hi ha dos tipus de valors atípics: els que són deguts als errors de mesura i els que provenen de les cues de les distribucions que generen les dades. Per tant, els valors atípics es poden considerar com a atípics segons quina siga la distribució que els modeli ja que, per exemple, si es suposa que les dades provenen d'una distribució normal, hi haurà tendència a interpretar els valors extrems com a valors atípics. En canvi, és possible que

aquests valors no siguin tant extrems en alguna altra distribució, per tant, una bona manera de combatre els valors atípics és utilitzar una distribució que els ajusti.

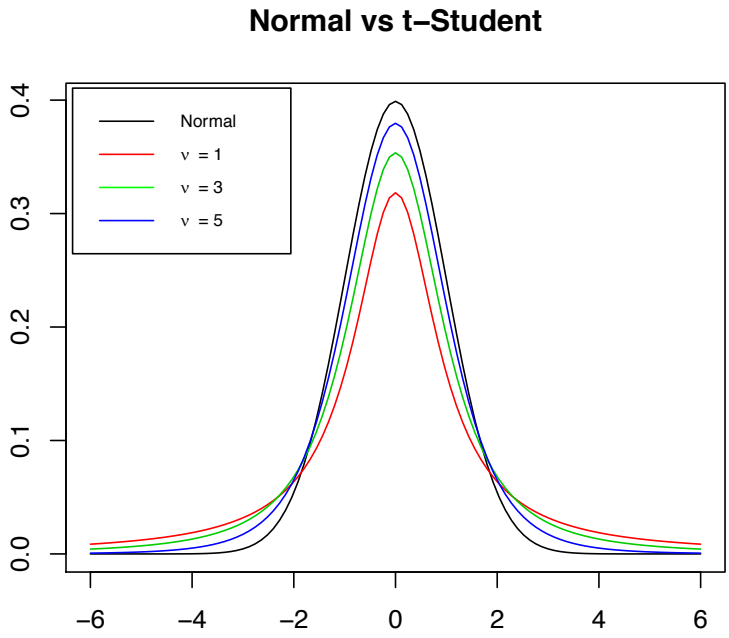
És sabut que la regressió lineal assumeix que el terme de pertorbació es distribueix com una normal amb mitjana $\mu = 0$ i variància constant igual a σ^2. Atès que aquesta distribució té les cues bastant estretes, aquest mètode classic de regressió és extramadament vulnerable a les dades extremes. La solució que proposa el mètode de la regresió robusta t a aquest problema és substituir la distribució normal del terme de pertorbació per una distribució amb cues més amples com és la distribució t-Student. És important ressaltar que la distribució t utilitzada a les proves d'hipòtesis convencionals no funciona ja que no es pot controlar la seva mitjana ni la seva variància, però es poden assignar fàcilment paràmetres de localització i escala si s'escriu la seva funció de densitat de la següent forma:

$$f(x;\mu,\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma}}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Aquesta parametrització s'anomena distribució t-Student no estandaritzada i permet controlar la mitjana (que és μ quan $\nu > 1$) i la variància (que és $\frac{\sigma^2\nu}{\nu-2}$ quan $\nu > 2$) tal com es pot fer amb la distribució normal, i també l'amplada de les cues mitjançant el paràmetre ν, que representa els graus de llibertat de la distribució i marca la influència que tindran els valors extrems. Cal destacar que els graus de llibertat de la distribució es poden determinar a priori o es poden estimar a partir de les dades de la mostra. La primera opció simplifica de forma considerable el mètode ja que el paràmetre es marca a priori i se li assigna el valor que l'investigador creu convenient per tal de donar-li un nivell de robustesa apropiat al model. Per contra, la segona opció és més sofisticada però aporta al model una robustesa adaptativa ja que es va calculant el valor del paràmetre a mesura que s'analitzen les dades de la mostra.

Seguidament s'adjunta en un gràfic la funció de densitat de la distribució normal (amb mitjana 0 i variància 1) i la de la distribució t-Student parametritzada amb diferents graus de llibertat per comparar-ne les amplades de les seves cues:

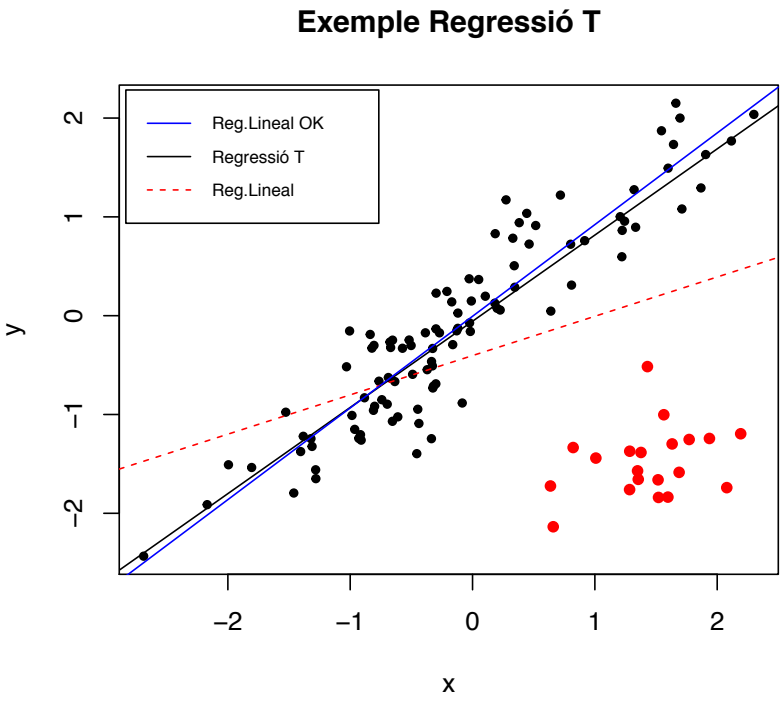
Imatge 5.3. Comparativa de la distribució normal i la t-Student segons els graus de llibertat.



Aquest gràfic és útil per comprovar que la distribució t-Student amb un grau de llibertat ($\nu = 1$) és la que té les cues més amples i que a mesura que s'augmenta el nombre de graus de llibertat de la distribució, les cues es van fent més estretes i la distribució s'assembla més a una normal amb mitjana zero i variància 1.

Per acabar aquest apartat s'inclourà a tall d'exemple en un gràfic el conjunt de dades utilitzat a l'apartat anterior (amb 100 dades "bones" i 20 de "dolentes" que es poden considerar com atípiques) i tal com s'ha fet anteriorment amb la regressió quantil lineal, apareixeran les tres rectes de regressió referents a la regressió lineal amb totes les dades, regressió lineal aplicada només al conjunt de 100 dades "bones" i regressió robusta t aplicada sobre totes les dades. D'aquesta manera es podrà comprovar de forma visual la robustesa d'aquest mètode que s'acaba d'exposar:

Imatge 5.4. Exemple d'aplicació de la regressió robusta T



Tal com s'ha dit abans, la robustesa d'aquest mètode és més que evident, ja que al veure el gràfic, la recta de regressió del mètode robust és quasi igual que la recta de regressió lineal aplicada sobre les dades que no tenen cap valor atípic. Per tant, la regressió robusta t, igual com la regressió quantil lineal, és una bona alternativa a la regressió lineal davant d'una possible situació en que es tinguin valors atípics a la mostra d'estudi.

5.3. REGRESSIÓ PER MÍNIMS QUADRATS PONDERATS

El tercer i últim mètode que s'analitzarà en aquest treball és el de la regressió per mínims quadrats ponderats i es podria dir que, generalment, és el més utilitzat en l'àmbit de l'estadística quan es vol dur a terme un mètode de regressió robusta. Aquest mètode consisteix, a grans trets, en definir una funció de pesos anomenada $w(\epsilon_i)$, que donarà un pes més petit a les observacions que tinguin un residu de la regressió ϵ_i "gran" per tal de reduir o inclús eliminar la influència que aquestes observacions tenen en la regressió.

Per introduir la funció que es minimitza a l'hora de calcular les estimacions dels paràmetres mitjançant els mínims quadrats ponderats, primer cal recordar quina és la que es minimitza per obtindre les estimacions per mínims quadrats ordinaris:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Per tant, la funció que cal minimitzar per obtenir les estimacions per mínims quadrats ponderats és:

$$S_p = \sum_{i=1}^{n} w(\epsilon_i)(\epsilon_i^2) = \sum_{i=1}^{n} w(y_i - \beta_0 - \beta_1 x_i)(y_i - \beta_0 - \beta_1 x_i)^2$$

on $w(\epsilon_i)$ és la funció de pesos que abans s'ha introduït, que té com a objectiu llevar-li influència a les observacions que tinguin els residus molt alts. Aquesta funció de pesos cal que compleixi les següents propietats:

- $w(\epsilon_i) \geq 0$
- $w(0) = 0$
- $w(\epsilon) = w(-\epsilon)$
- Si $|\epsilon_i| < |\epsilon_j|$, llavors $w(\epsilon_i) \leq w(\epsilon_j)$

Aquesta funció de pesos $w(\epsilon_i)$ pot estar definida de diverses formes. Algunes de les propostes són la de Hampel (introduïda per Hampel l'any 1968), la del sinus (introduïda per Andrews l'any 1974) o la biquadarda (introduida per Beaton i Tukey l'any 1974). Però la més utilitzada és la de Huber, la qual ve donada per la següent expressió:

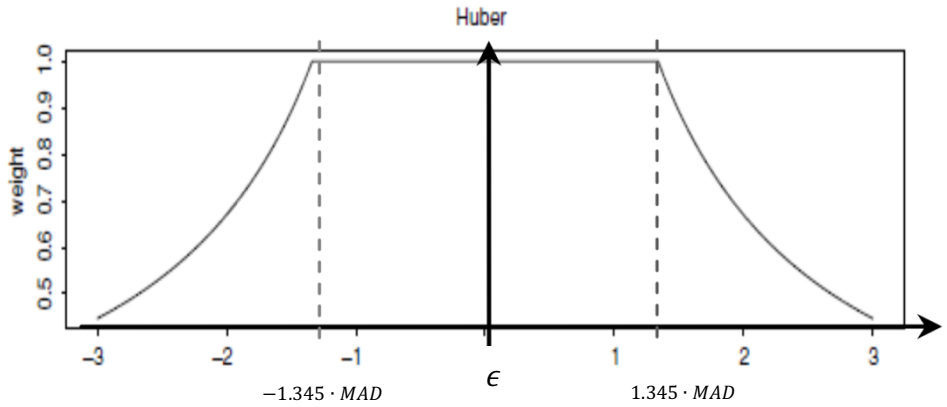
$$w(\epsilon) = \begin{cases} 1, & |\epsilon| \leq k \\ \dfrac{k}{|\epsilon|}, & |\epsilon| > k \end{cases}$$

on k s'anomena constant d'afinat (en anglès *tunning constant*) i quan els residus són normals, es fixa al valor 1.345σ (on σ és la desviació estàndard dels residus) ja que aquest valor proporciona una eficiència del mètode del 95% i a part produeix una resistència acceptable en front dels valors atípics. Tanmateix, però, hi ha una forma de mesurar la variabilitat dels residus que és molt més robusta que la desviació estàndard. Aquesta és la Desviació Absoluta de la Mediana (MAD) que es calcula:

$$MAD = mediana(|\epsilon_i - mediana(\epsilon)|)$$

A la pràctica, quan s'utilitza el mètode dels mínims quadrats ponderats amb la funció de pesos de Huber, és aquesta la mesura de dispersió utiltizada per calcular la constant d'afinat k. La funció de pesos definida per Huber té aquesta forma:

Imatge 5.5. Funció de pesos definida segons Huber.

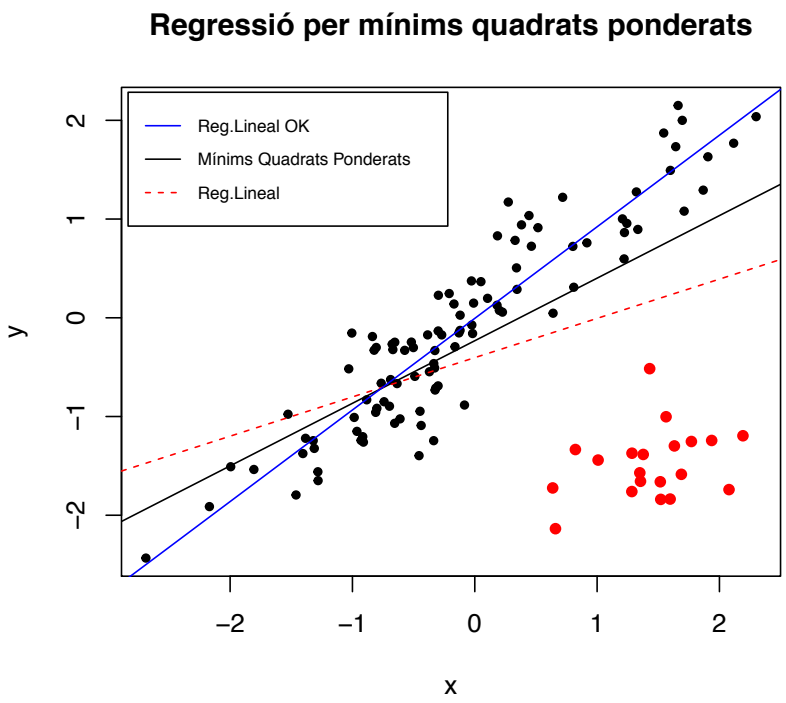


A partir del gràfic de la funció de pesos de Huber es pot veure com els residus que es troben acotats dins del rang de la constant d'afinat en negatiu i en positiu tenen un pes d'1 i els que no, tenen un pes inferior, ja que a mesura que s'augmenta el valor del residu (tant en positiu com en negatiu) el pes d'aquest és veu disminuït.

Un cop s'ha definit la funció objectiu que minimitza el mètode i la funció de pesos que utilitza per tal de restar-li importància a les observacions amb residus molt grans és hora de presentar el funcionament del mètode, i la forma en que aquest arriba a les estimacions finals dels paràmetres. En primer lloc cal destacar que el mètode dels mínims quadrats ponderats és un mètode iteratiu. Aquest mètode, d'igual forma que tots els mètodes iteratius necessita uns valors d'entrada per començar a iterar, els quals en aquest cas seran els paràmetres i residus obtinguts per l'ajust d'un model de regressió lineal per mínims quadrats ordinaris, els quals, tots i cadascun d'ells tindran un pes d'1. A partir d'aquests paràmetres, residus i pesos inicials, es van calculant de forma iterativa nous paràmetres,residus i pesos per tal de minimitzar la funció objectiu del mètode fins arribar a la convergència del procés, o sigui, fins que els valors dels paràmetres, residus i pesos canvien molt poc d'una iteració a una altra.

De la mateixa forma que s'ha fet amb els altres dos mètodes de regressió robusta exposats, per tal de demostrar gràficament l'eficiència d'aquest mètode, s'utilitzaran les dades amb valors atípics que s'han creat en anterioritat per comparar la regressió per mínims quadrats ponderats amb la regressió lineal que només té en compte les dades "bones" i també amb la regressió lineal amb totes les dades. D'aquesta forma es veurà gràficament com afecten les dades atípiques a la regressió lineal per mínims quadrats ordinaris i també la robustesa que tenen els mínims quadrats ponderats en front d'aquest tipus d'observacions que violen la hipòtesi de normalitat.

Imatge 5.6. Exemple d'aplicació de la regressió per mínims quadrats ponderats.



A partir del gràfic es pot comprovar que la recta de regressió calculada per mínims quadrats ponderats no es veu tant afectada pels valors atípics com la recta de regressió calculada per mínims quadrats ordinaris, per tant, òbviament, la primera té una major robustesa que la segona. Tanmateix, si es compara aquesta recta de regressió robusta amb les dues que s'han calculat als dos capítols anteriors (regressió quantil lineal i regressió robusta T) es pot comprovar que és la dels mínims quadrats ponderats la que presenta una menor robustesa ja que és la que es veu més desviada respecte la recta de regressió lineal en que només es tenen en compte les dades "bones".

6. ANÀLISI PRÀCTIC AMB R

Aquest darrer punt del treball consisteix en realitzar l'anàlisi d'un conjunt de dades reals mitjançant el software R. Després d'analitzar en detall la part teòrica i el funcionament de la regressió lineal i dels mètodes de regressió robusta exposats, és torn d'endinsar-se en un problema més pràctic mitjançant una base de dades amb una variable resposta i diverses variables regressores, i dur a terme un anàlisi de regressió lineal múltiple, el qual s'aproxima més a un estudi real de mètodes predictius que no pas els petits exemples exposats anteriorment de regressió lineal simple en que només apareixia una sola variable regressora.

6.1. DESCRIPCIÓ DE LA BASE DE DADES

La base de dades que s'utilitzarà per a dur a terme l'anàlisi pràctic amb R està composada per 5875 mesures de veu biomèdiques de 42 persones malaltes de *Parkinson* en etapa prematura (hi ha moltes observacions per pacient), les quals van estar reclutades per a un assaig d'un dispositiu de telemonitorització per al seguiment remot de la progressió dels psíntomes durant sis mesos. Aquestes grabacions de veu van estar grabades de forma automàtica des de la casa de cada pacient. Cal destacar que la base de dades s'ha extret del *Machine Learning Repository* i que es pot trobar al següent link:

http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

Les mesures de veu biomèdiques dels pacients són 16, però a part, encara que no s'utilitzaran a l'estudi, a la base de dades també hi apareix l'edat de cada pacient, el sexe i el temps en dies des de la data de reclutament. Per altra banda, la variable d'interès és una variable numèrica que s'anomenta total_UPDRS (de l'anglés, *Unified Parkinson's Disease Rating Scale*) i és una mesura majorment acceptada dins de la comunitat científica per valorar la gravetat de l'enfermetat de *Parkinson* en una persona. Convé destacar que hi ha una darrera variable que s'anomena motor_UPDRS, que fa referència a una mesura d' UPDRS similar a la variable d'interès, que no serà inclosa en aquest estudi.

És sabut que les 16 mesures de veu biomèdiques són independents i afecten a la puntuació total_UPDRS, per tant, l'objectiu de l'estudi serà utilitzar-les per predir aquesta puntuació mitjançant tècniques de regressió. A continuació s'adjunta una petita descripció d'aquestes 16 mesures:

- Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5 i Jitter:DDP mesuren la variació de la freqüència fonamental de la veu.

- Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11 i Shimmer:DDA mesuren la variació en l'amplitud de la veu.

- NHR,HNR mesuren la relació entre el soroll i els components tonals de la veu.

- RPDE és una mesura no lineal de complexitat dinàmica.

- DFA és un index escalat a nivell fractal.

- PPE és una mesura no lineal de la variació de la freqüència fonamental.

6.2. IMPORTACIÓ I PREPROCESSAMENT DE LES DADES

En primer lloc, com que les dades es troben en un fitxer ".txt" s'ha fet servir la funció "read_csv" del paquet "readr" per tal d'importar-les. El data frame on aquestes dades s'han guardat a la memòria d'R s'ha anomenat "dd":

```
library(readr)
dd <- read_csv("~/Desktop/DADES TFG/parkinson.txt")
```

Tot seguit s'ha comprovat la classe de cada variable amb la funció "sapply", ja que és important assegurar-se abans de continuar amb l'estudi que cada variable està definida amb el format que li pertoca, vist que si no és així es pot arribar a resultats invàlids. Aquesta funció consisteix en calcular qualsevol altra funció, que en aquest cas serà la funció "class", per a totes i cadascuna de les columnes d'un data frame:

```
sapply(dd,class)
```

```
##      subject#           age           sex     test_time   motor_UPDRS
##     "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
##   total_UPDRS     Jitter(%)   Jitter(Abs)    Jitter:RAP   Jitter:PPQ5
##     "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
##    Jitter:DDP       Shimmer   Shimmer(dB)  Shimmer:APQ3  Shimmer:APQ5
##     "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
## Shimmer:APQ11   Shimmer:DDA           NHR           HNR          RPDE
##     "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
##           DFA           PPE
##     "numeric"     "numeric"
```

Vist aquest resultat, les 16 variables numèriques que fan referència a mesures de veu biomèdiques estan ben codificades ja que la classe de cadascuna és numèrica. En canvi, no es pot dir el mateix de l'identificador de cada pacient, ja que la variable "subject#" està codificada com a numèrica i caldrà transformar-la a factor (també es podria recodificar com a factor la variable "sex" però com no es farà servir a l'estudi no es farà):

```
dd$`subject#` <- as.factor(dd$`subject#`)
```

A partir de la sortida de la funció "sapply" s'ha comprovat que els noms de les variables de la base de dades no són molt convincents ja que alguns contenen símbols que no són gens recomanats a l'hora de treballar amb noms d'elements a R. Aquests símbols són els dos punts, els parèntesis, el símbol del percentatge, etc... Per tant, es durà a terme una recodificació de tots aquests noms de la següent forma:

```
names(dd) <- c("pacient","age","sex","test_time",
               "motor_UPDRS","total_UPDRS","jitter_pct",
               "jitter_abs","jitter_RAP","jitter_PPQ5",
               "jitter_DDP","shimmer","shimmer_db",
               "shimmer_APQ3","shimmer_APQ5","shimmer_APQ11",
               "shimmer_DDA","NHR","HNR","RPDE","DFA","PPE")
```

Després de donar-li a les variables noms més usuals, sense cap tipus de símbol (com els parèntesis o el símbol del percentatge), es crearà un nou conjunt de dades que conté només les variables que s'utilitzaran més endavant en els models de regressió. Aquestes són l'indicador de cada pacient, la variable resposta (total_UPDRS) i les 16 variables numèriques que hi ha darrere d'aquesta. Aquest nou conjunt de dades serà també un data frame i s'anomenarà "dd2":

```
dd2 <- dd[,c(1,6:22)] #Només les variables a introduir al model
```

Ara que ja es té el conjunt de dades que s'utilitzarà per construir els models de regressió és moment de fer una descriptiva numèrica univariant de cada variable per tenir una primera visualització del comportamnet de les dades i comprovar si hi ha alguna variable que conté dades mancants o *missings*. Per fer això s'utilitzarà un altre cop la funció "sapply", però ara, en contes de calcular la classe de cada variable, es calcularà el resum numèric (mitjançant la funció "summary" d'R):

```
sapply(dd2,summary)
```

```
## $pacient
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
## 149 145 144 137 156 156 161 150 152 148 138 107 112 136 143 138 144 126
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
```

```
## 129 134 123 112 138 156 144 130 129 134 168 126 130 101 135 161 165 129
##  37  38  39  40  41  42
## 140 149 143 142 165 150
##
## $total_UPDRS
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   21.37   27.58   29.02   36.40   54.99
##
## $jitter_pct
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000830 0.003580 0.004900 0.006154 0.006800 0.099990
##
## $jitter_abs
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 2.250e-06 2.244e-05 3.453e-05 4.403e-05 5.333e-05 4.456e-04
##
## $jitter_RAP
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000330 0.001580 0.002250 0.002987 0.003290 0.057540
##
## $jitter_PPQ5
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000430 0.001820 0.002490 0.003277 0.003460 0.069560
##
## $jitter_DDP
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000980 0.004730 0.006750 0.008962 0.009870 0.172630
##
## $shimmer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00306 0.01912 0.02751 0.03404 0.03975 0.26863
##
## $shimmer_db
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.026   0.175   0.253   0.311   0.365   2.107
##
## $shimmer_APQ3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00161 0.00928 0.01370 0.01716 0.02057 0.16267
##
## $shimmer_APQ5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00194 0.01079 0.01594 0.02014 0.02375 0.16702
##
## $shimmer_APQ11
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00249 0.01566 0.02271 0.02748 0.03272 0.27546
##
## $shimmer_DDA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00484 0.02783 0.04111 0.05147 0.06173 0.48802
##
## $NHR
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

```
## 0.000286 0.010955 0.018448 0.032120 0.031463 0.748260
##
## $HNR
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.659  19.406  21.920  21.680  24.444  37.875
##
## $RPDE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1510  0.4698  0.5423  0.5415  0.6140  0.9661
##
## $DFA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5140  0.5962  0.6436  0.6532  0.7113  0.8656
##
## $PPE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02198 0.15634 0.20550 0.21959 0.26449 0.73173
```

La descriptiva numèrica univariant de cada variable ens mostra que no hi ha cap dada mancant o *missing* en tota la base de dades ja que en cap d'aquests resums numèrics s'observa algun valor *NA.* A part d'això, la descriptiva univariant també mostra que hi ha algunes variables que podrien contindre valors atípics ja que la diferència entre el valor que marca el tercer quartil de la distribució de la variable i el valor màxim d'aquesta, és bastant gran. Alguns exemples són les variables NHR, shimmer_DDA o jitter_DDP.

Vist tot això, ja s'ha dut a terme la importació de les dades, el seu preprocessament i les primeres descriptives. Per tant, ja estan preparades per a dur a terme els models de regressió que es construiran als següents apartats.

6.3. MODEL DE REGRESSIÓ LINEAL

6.3.1. MODEL DE REGRESSIÓ LINEAL COMPLET

El primer model que cal construir és el de regressió lineal amb l'indicador de cada pacient que farà la funció de bloc, ja que es té més d'una observació per pacient, i totes les variables numèriques referents a les diferents mesures de veu biomèdiques com a regressores. Aquest model, anomenat model 1, és el següent:

```
model1 <- lm(total_UPDRS ~. , data=dd2)

summary(model1)

##
## Call:
```

```
## lm(formula = total_UPDRS ~ pacient + jitter_pct + jitter_abs +
##     jitter_RAP + jitter_PPQ5 + jitter_DDP + shimmer + shimmer_db +
##     shimmer_APQ3 + shimmer_APQ5 + shimmer_APQ11 + shimmer_DDA +
##     NHR + HNR + RPDE + DFA + PPE, data = dd2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2519 -1.6673  0.1083  1.6333 10.2458
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.997e+01  1.189e+00  33.627  < 2e-16 ***
## pacient2      -2.391e+01  3.927e-01 -60.883  < 2e-16 ***
## pacient3      -7.429e+00  3.265e-01 -22.751  < 2e-16 ***
   ...
   ...
## pacient41      2.319e+00  3.483e-01   6.657 3.05e-11 ***
## pacient42     -7.322e+00  3.312e-01 -22.109  < 2e-16 ***
## jitter_pct    -2.099e+02  6.615e+01  -3.173 0.001515 **
## jitter_abs     8.952e+03  3.314e+03   2.701 0.006935 **
## jitter_RAP     1.545e+03  1.333e+04   0.116 0.907732
## jitter_PPQ5    2.056e+02  5.800e+01   3.544 0.000397 ***
## jitter_DDP    -4.652e+02  4.444e+03  -0.105 0.916630
## shimmer        1.778e+01  1.896e+01   0.938 0.348337
## shimmer_db     1.293e+00  1.424e+00   0.908 0.363921
## shimmer_APQ3  -3.267e+03  1.338e+04  -0.244 0.807060
## shimmer_APQ5  -3.022e+01  1.607e+01  -1.881 0.060019 .
## shimmer_APQ11 -3.781e+00  7.284e+00  -0.519 0.603756
## shimmer_DDA    1.083e+03  4.459e+03   0.243 0.808088
## NHR           -3.587e+00  1.957e+00  -1.833 0.066843 .
## HNR            6.772e-02  2.520e-02   2.687 0.007233 **
## RPDE          -1.071e-01  6.403e-01  -0.167 0.867178
## DFA           -1.763e+00  1.185e+00  -1.488 0.136725
## PPE            4.260e-01  9.525e-01   0.447 0.654698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.761 on 5817 degrees of freedom
## Multiple R-squared:  0.9341, Adjusted R-squared:  0.9334
## F-statistic:  1446 on 57 and 5817 DF,  p-value: < 2.2e-16
```

El resum numèric del primer model és aquest que s'acaba d'adjuntar. Cal destacar que s'ha retallat un amica ja que, com que la variable indicadora de cada pacient és un factor amb 42 nivells, la sortida d'R del resum del model mostra per pantalla cadascun dels 42 nivells, cosa que ocupa molt d'espai.

El primer que es pot veure a partir d'aquest resum és que el test de significació global del model (test F de la darrera línea del resum) té un p.valor associat molt inferior al nivell de significació del 5%, per tant, es pot afirmar que el model és util per explicar la variable resposta ja que es rebutja la hipòtesi nul.la del test, la qual defensa que no hi ha cap variable explicativa

rellevant per explicar la variable resposta. Tot seguit, mitjançant l' R^2 i l' R^2 ajustat (0.9341 i 0.9334 respectivament) es pot comprovar que la bondat de l'ajust del model és força alta, ja que mitjançant la regressió s'explica aproximadament el 93% de la variabilitat total de la variable resposta, per tant, es pot afirmar que el model és força bo. Tot i això, hi ha bastantes variables explicatives que tenen un p.valor associat al test de significació individual de cada variable superior al nivell de significació del 5%, la qual cosa fa pensar que aquestes variables no són rellevants per explicar la variable resposta. Aquest problema es pot solucionar construint un nou model mitjançant un procés automàtic de selecció de variables, el qual només inclourà les variables explicatives que realment siguin rellevants per explicar la variable resposta.

6.3.2. MODEL DE REGRESSIÓ LINEAL AMB PROCÉS DE SELECCIÓ DE VARIABLES

El procés de selecció de variables és un mètode iteratiu que consisteix en comprovar, a cada iteració, si una variable forma part del model o no. Aquest procés es pot dur a terme de tres formes diferents:

- Cap enrere (*backward*): Es parteix del model complet i a cada iteració s'elimina la variable menys rellevant tenint en compte el contrast de significació individual de cadascuna. El procés acaba quan totes les variables que queden al model són rellevants.

- Cap endavant (*fordward*): Es parteix d'un model nul (sense cap variable explicativa) i s'introdueixen les variables regressores de forma seqüencial. La primera variable explicativa en entrar al model és la que té una major correlació (positiva o negativa) amb la variable resposta i entra si compleix el criteri d'entrada al model, i després, es consideren les que tenen una major correlació parcial amb aquesta variable resposta. El procés acaba quan no queden variables que compleixin els criteris d'entrada al model.

- Cap endavant i cap enrere (*stepwise*): Combinació dels dos mètodes anteriors, però que té com a diferència que en cada iteració comprova si cal introduir alguna variable prèviament exclosa o si per contra, cal eliminar alguna variable prèviament introduïda. Es pot partir del model nul o del model complet.

En aquest treball s'utilitzarà el darrer mètode exposat, *l'stepwise,* ja que és el més efectiu a l'hora de crear un model de regressió pel fet de tenir en compte totes les variables a totes les iteracions. Aquest procés de selecció de variables es durà a terme amb l'R utilitzant la funció "step".

A la descripció del mètode *fordward* de selecció de variables s'ha explicat que per tal d'introduir una variable al model, aquesta cal que compleixi el criteri d'entrada. Aquest criteri és un dels paràmetres (paràmetre *k*) que cal definir a la funció "step" de l'R. Hi ha dos criteris que poden utilitzar-se amb aquesta funció:

- AIC: criteri d'informació d'*Akaike* (de l'anglès, *Akaike Information Criterion*) que mesura la bondat de l'ajust o la qualitat relativa d'un model estadístic. Serveix per seleccionar entre diferents models, ja que quan més petit és aquest valor, millor ajust té el model. Es calcula:

$$AIC = 2k - 2\ln(L)$$

  on *k* és el nombre de paràmetres que té el model i *L*, el valor màxim de la funció de versemblança per al model donat.

- BIC: criteri d'informació bayesiana (de l'anglès, *Bayesian Information Criterion*) que serveix per seleccionar entre diferents models estadístics. Està estrictament relacionat amb l'AIC però es diferencia en que és més restrictiu a l'hora d'incorporar variables al model ja que té un terme de penalització d'introducció de paràmetres al model superior al que té l'AIC. Es calcula:

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$$

  on *k* i *L*, com s'ha explicat abans, són el nombre de paràmetres i el valor màxim de la funció de versemblança respectivament, i *n*, el nombre de dades disponibles per construir el model estadístic, o dit d'una altra forma, la mida mostral.

Tal i com s'ha avançat a la descripció del criteri BIC, aquest és més restrictiu que l'AIC i utilitzant-lo, s'obtindran models generalment amb un menor nombre de variables. Per tant, els models obtinguts amb el criteri BIC seran més senzills i faran prediccions menys detallades que si s'utilitza el criteri AIC. Cal destacar, que després de valorar les dues opcions, s'ha decidit d'utilitzar el criteri AIC per construir un nou model a partir d'un procés automàtic de selecció de variables ja que el model que s'obté utilitzant el criteri BIC està format per només una variable explicativa, cosa que no interessa en aquest estudi, ja que es vol dur a terme un anàlisi de regressió lineal múltiple amb diverses variables regressores. El model obtingut ha sigut el següent:

```
model_AIC <- step(model1,direction="both",k = 2,trace = 0)
summary(model_AIC)
```

```
##
## Call:
```

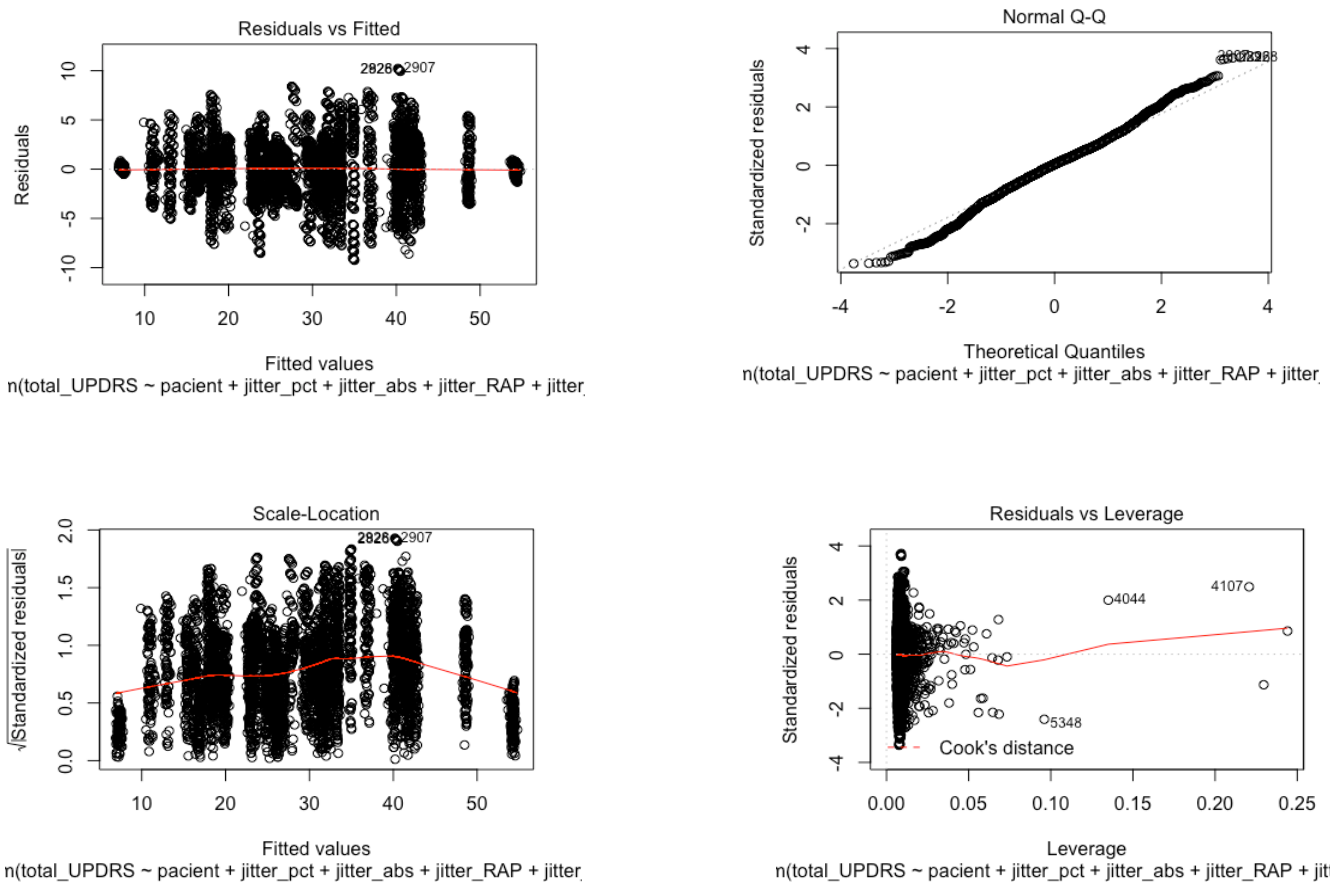
```
## lm(formula = total_UPDRS ~ pacient + jitter_pct + jitter_abs +
##     jitter_RAP + jitter_PPQ5 + shimmer_db + shimmer_APQ5 + NHR +
##     HNR + DFA, data = dd2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2550 -1.6649  0.1132  1.6304 10.2266
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    39.95049    1.00004  39.949  < 2e-16 ***
## pacient2      -23.93859    0.38730 -61.809  < 2e-16 ***
## pacient3       -7.43953    0.32474 -22.909  < 2e-16 ***
   ...
   ...
## pacient41       2.33252    0.34401   6.780 1.32e-11 ***
## pacient42      -7.33529    0.32568 -22.523  < 2e-16 ***
## jitter_pct   -200.72346   61.53716  -3.262 0.001113 **
## jitter_abs   9096.52858 3136.54431   2.900 0.003743 **
## jitter_RAP    131.58743   78.15105   1.684 0.092282 .
## jitter_PPQ5   212.01258   56.64371   3.743 0.000184 ***
## shimmer_db      1.91374    0.85348   2.242 0.024980 *
## shimmer_APQ5  -30.95600   11.58617  -2.672 0.007565 **
## NHR            -3.29548    1.87438  -1.758 0.078771 .
## HNR             0.06684    0.02188   3.055 0.002263 **
## DFA            -1.66757    1.17272  -1.422 0.155089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.759 on 5824 degrees of freedom
## Multiple R-squared:  0.9341, Adjusted R-squared:  0.9335
## F-statistic:  1650 on 50 and 5824 DF,  p-value: < 2.2e-16
```

En primer lloc, cal destacar que de les 16 variables numèriques referents a mesures de veu biomèdiques, després d'aplicar el procés automàtic de selecció de variables utilitzant el criteri AIC, han quedat només 9. Es pot veure, però, que el test de significació individual associat a cada variable és, o inferior al nivell de significació del 5%, o una mica superior però no major que 0.15, cosa que no passava anteriorment amb el model complet ja que hi havia certes variables explicatives que tenien un p.valor associat al test de significació individual de l'ordre de 0.9. Abans de dur a terme la validació i la interpretació d'aquest model, cal destacar que la funció "step" de l'R ha estat implementada amb els paràmetres *direction = "both"* (procés de selecció *stepwise*), *k = 2* (criteri AIC, ja que si es vol utilitzar el criteri BIC cal donar-li al paràmetre *k* el valor *log(nrow(dd2))* i *trace = 0*, aquest últim per tal que l'R no mostri per pantalla totes i cadascuna de les iteracions que realitza el procés automàtic de selecció de variables.

Seguidament es validarà el model mitjançant el resum numèric del mateix que s'acaba d'adjuntar i els gràfics dels seus residus. En primer lloc, el test de significació global té un

p.valor associat molt inferior al nivell de significació del 5%, per tant el model és valid. A part d'això, l'R^2 ajustat ha augmentat lleugerament respecte el model complet (en aquest cas és 0.9335 i el del model complet, 0.9334) com a conseqüència d'eliminar certes variables que no aportaven pràcticament res a la regressió, per tant, es pot afirmar que el model s'ajusta molt bé i que explica un percentatge molt elevat (93.35%) de la variabilitat de la variable resposta. Tal com s'ha explicat abans, el test de significació individual de les variables explicatives és força petit en cadascuna d'elles, sent el d'algunes variables bastant inferior al nivell de significació del 5% i el d'algunes altres, una mica superior. Això útlim és degut a que s'ha utilitzat el criteri AIC per construir el model mitjançant el procés automàtic de selecció de variables, el qual no és tant restrictiu com el criteri BIC, ja que si s'haguera fet servir aquest últim, el model només inclouria les variables amb p.valors associats als test de significació individual molt inferiors al nivell de significació del 5%. Arribat aquest punt, si un només es fixa en el resum numèric del model, aquest podria ser validat, però això no és suficient ja que cal fixar-se amb els gràfics dels residus, els quals s'adjunten a continuació:

Imatge 6.1. Gràfics dels residus del model de regressió lineal (model_AIC).



El primer gràfic fa referència als residus de la regressió (eix Y) en front dels valors predits per aquesta (eix X). Aquest gràfic mostra uns residus que estan distribuïts de forma força aleatòria al voltant del valor 0 que no presenten cap tipus de tendència ni forma anòmala que pugui fer dubtar sobre la validació del model. Tanmateix, es pot apreciar que apareixen moltes observacions molt mal predites pel model ja que tenen residus que arriben fins al valor 10 en positiu i en negatiu, la qual cosa és una mostra de la presència de valors atípics al conjunt de dades. Per tant, amb aquest primer gràfic, per una part es podria validar el model ja que els residus prenen una tendència lineal al voltant del zero, però per una altra no perque la presència de molts valors atípics podria distorsionar molt les estimacions dels paràmetres i per tant, les prediccions podrien no ser les més fiables.

El segon gràfic, que mostra els residus estudentitzats (eix Y) en front dels quantils teòrics (eix X) en paper probabilístic normal (el que s'anomena en estadística *QQ-Plot*), permet apreciar que la normalitat dels residus pateix sobretot a les cues, la qual cosa podria estar causada per la presència dels valors atípics que s'acaba de comentar al paràgraf anterior.

El tercer gràfic permet concluir el mateix que el primer perque mostra el mateix però en contes de plasmar els residus a l'eix Y, mostra l'arrel quadrada dels residus estudentitzats.

I finalment, el quart gràfic mostra les observacions que poden ser més influents del conjunt de dades ja que a l'eix Y es troben els residus estudentitzats i a l'X, el *Leverage*, que és una mesura que quantifica com de lluny es troba una observació respecte les altres sense tenir en compte el valor de la variable resposta. Destacar que una observació pot tenir un alt valor de *Leverage* sense tenir una gran influència en la regressió, ja que encara que estigui lluny de les altres observacions, pot estar prop de la recta de regressió, o sigui, tenir un residu petit. Aquest fet es pot comprovar amb el quart gràfic de la validació del model ja que hi ha dos observacions que tenen un Leverage molt alt (quasi 0.25) però per altra banda tenen un residu acceptable ja que totes dos no superen el 2 ni el -2 respectivament. Per altra banda, les observacions 4044, 4107 i 5348 són les que tenen gran influència en la regressió ja que a part d'estar allunyades de la resta d'observacions (alt Leverage), també tenen un alt valor de residu, per tant aquests són el tipus d'observacions que distorsionen més les prediccions dels models de regressió.

Englobant les interpretacions extretes a partir del resum numèric del model i dels gràfics d'aquest, es pot arribar a la conclusió de que el model és vàlid i s'ajusta bé a les dades però que la presència de valors atípics, pot causar que les estimacions dels paràmetres no siguin del tot correctes. Per tant, com que el treball tracta sobre els mètodes de regressió robusta, s'aplicaran aquests mètodes i es comprovarà si les seves estimacions, les quals no estaran distorsionades per la presència dels atípics, coincideixen o no, amb les del model de regressió lineal.

6.4. MODEL DE REGRESSIÓ QUANTIL LINEAL

En aquest apartat es construirà un model de regressió quantil lineal a partir del que s'ha obtingut en l'apartat anterior després d'aplicar el procés automàtic de selecció de variables. Tal i com s'ha avançat anteriorment en l'apartat en que s'ha introduït la regressió quantil lineal, s'utilitzarà la funció "rq" de la llibreria "quantreg", i el valor del paràmetre "tau" serà el que ve per defecte amb la funció, o sigui 0.5, el qual estima la mediana de la variable resposta. Aquest model és el següent:

```
library(quantreg)

m2_q <- rq(total_UPDRS ~ pacient + jitter_pct +
             jitter_abs + jitter_RAP + jitter_PPQ5 +
             shimmer_db + shimmer_APQ5 + NHR + HNR + DFA,data = dd2)

summary(m2_q)

## Warning in summary.rq(m2_q): 19 non-positive fis

##
## Call: rq(formula = total_UPDRS ~ pacient + jitter_pct + jitter_abs +
##     jitter_RAP + jitter_PPQ5 + shimmer_db + shimmer_APQ5 + NHR +
##     HNR + DFA, data = dd2)
##
## tau: [1] 0.5
##
## Coefficients:
##               Value      Std. Error t value    Pr(>|t|)
## (Intercept)    41.75509    0.72668   57.45981    0.00000
## pacient2      -24.86232    0.57406  -43.30995    0.00000
## pacient3       -6.97572    0.96481   -7.23012    0.00000
...

...
## pacient41       3.14605    0.59366    5.29942    0.00000
## pacient42      -6.98522    0.59982  -11.64547    0.00000
## jitter_pct    -81.94828   57.67415   -1.42088    0.15540
## jitter_abs   4102.97550 3330.21774    1.23204    0.21798
## jitter_RAP      9.46175   54.24357    0.17443    0.86153
## jitter_PPQ5   142.60897   45.86257    3.10948    0.00188
## shimmer_db      0.66904    0.56502    1.18409    0.23643
## shimmer_APQ5  -16.59593    8.87251   -1.87049    0.06147
## NHR            -1.03298    1.24770   -0.82791    0.40776
## HNR             0.02004    0.01028    1.94867    0.05138
## DFA            -3.14349    0.74369   -4.22691    0.00002
```

Després d'observar la sortida del model de regressió quantil lineal, el que més crida l'atenció és el valor de les estimacions dels seus paràmetres. La major diferència entre aquests paràmetres i els que s'han estimat mitjançant el model de regressió lineal, és que els de la

regressió quantil lineal són més petits en valor absolut, o sigui, tant els positius com els negatius són més propers a zero (excepte el paràmetre associat a la variable "DFA" que val -1.67 al model de regressió lineal i -3.14 al model de regressió quantil lineal). Tot seguit s'adjunten en una taula aquestes estimacions per poder comprovar-ho de forma més clara:

Taula 6.1. Comparativa dels coeficients de la regressió lineal i la regressió quantil lineal.

	Regressió Lineal	**Regressió Quantil Lineal**
Jitter_pct	-200.72	-81.95
Jitter_abs	9096.53	4102.97
Jitter_RAP	131.58	9.46
Jitter_PPQ5	212.01	142.61
Shimmer_db	1.91	0.67
Shimmer_APQ5	-30.95	-16.59
NHR	-3.29	-1.03
HNR	0.06	0.02
DFA	-1.67	-3.14

Aquestes diferències que s'acaben d'exposar entre les estimacions dels paràmetres del model de regressió quantil lineal i els del model de regressió lineal són del tot coherents, ja que la presència d'observacions atípiques a la mostra d'estudi provoca que les estimacions dels paràmetres del model de regressió lineal siguin més extremes que les del model de regressió quantil lineal com a conseqüència de la major vulnerabilitat que té el model de regressió lineal en front d'aquestes dades que violen la hipòtesis de normalitat.

Per tant, les prediccions de la variable resposta "total_UPDRS" que es poden fer a partir del model de regressió quantil lineal seran, generalment, més moderades (petites) que les que s'extrauran a partir del model de regressió lineal, ja que les aportacions sobre la variable resposta que tenen les variables explicatives són menors al tindre unes estimacions més petites en valor absolut.

6.5. MODEL DE REGRESSIÓ ROBUSTA T

Tot seguit es construirà el model de regressió robusta T amb les mateixes variables explicatives que s'han utilitzat en l'apartat anterior, o sigui, l'indicador del bloc de cada pacient i les 9 variables numèriques que s'han obtingut després de realitzar el procés automàtic de selecció de variables. La peculiaritat que es troba en aquest apartat és que la funció "treg" del paquet "SMIR" no funciona del tot bé. Un cop construït el model, quan es vol extraure el resum numèric (*summary*), aquest paquet extrau els mateixos coeficients que s'extrauen amb la regressió lineal. El que cal fer per tal d'obtenir els coeficients estimats després de realitzar la regressió robusta T és extraure els "tcoef" del model. Destacar, que el paràmetre "r" de la

funció “treg” fa referència als graus de llibertat de la distribució t-Student amb que es treballarà, els quals seran 1 ja que, com s’ha vist anteriorment, és el nombre de graus de llibertat amb els quals s’obté una distribució amb les cues més amples, cosa que provoca una major robustesa. Abans de mostrar els paràmetres estimats amb aquest mètode de regressió robusta cal destacar que el paquet “SMIR” ha sigut retirat de les llibreries d’R com a conseqüènca dels errors referents al resum numèric del model que s’acaben de descriure.

```
library(SMIR)

m4_t <- treg(model_AIC, r = 1.1, verbose = F)
m4_t$tcoef

##   (Intercept)      pacient2      pacient3         ...         ...
##  4.031189e+01 -2.453117e+01 -4.958652e+00
##     pacient41     pacient42    jitter_pct    jitter_abs    jitter_RAP
##  3.860790e+00 -6.305738e+00  6.030680e+01 -6.152149e+03 -9.269421e+00
##   jitter_PPQ5    shimmer_db  shimmer_APQ5           NHR           HNR
##  2.563410e+00 -1.025564e-01 -3.364253e+00 -5.305959e-02  5.933564e-03
##           DFA
## -1.223036e+00
```

Taula 6.2. Comparativa dels coeficients de la regressió lineal i la regressió robusta T.

	Regressió Lineal	**Regressió Robusta T**
Jitter_pct	-200.72	60.3068
Jitter_abs	9096.53	-6152.149
Jitter_RAP	131.58	-9.269421
Jitter_PPQ5	212.01	2.56341
Shimmer_db	1.91	-0.1025564
Shimmer_APQ5	-30.95	-3.364253
NHR	-3.29	-0.05305959
HNR	0.06	0.005933564
DFA	-1.67	-1.223036

Un cop obtingudes les estimacions dels paràmetres del model de regressió robusta T es pot comprovar que, generalment, són més petites que les del model de regressió lineal (tal com passava abans amb el model de regressió quantil) i que en alguns cops, les estimacions fins i tot canvien de signe.

6.6. MODEL DE REGRESSIO PER MÍNIMS QUADRATS PONDERATS

Finalment, en aquest darrer apartat es construirà el model de regressió per mínims quadrats ponderats utilitzant com a variables explicatives les que s’han obtingut mitjançant el procés automàtic de selecció de variables amb el criteri de l’AIC. Per construir-lo s’utilitzarà la funció “rlm” de la llibreria “MASS”. El model és el següent:

```r
library(MASS)

m3_w <- rlm(total_UPDRS ~ pacient + jitter_pct +
             jitter_abs + jitter_RAP + jitter_PPQ5 +
             shimmer_db + shimmer_APQ5 + NHR + HNR + DFA,data = dd2)


summary(m3_w)

##
## Call: rlm(formula = total_UPDRS ~ pacient + jitter_pct + jitter_abs +
##     jitter_RAP + jitter_PPQ5 + shimmer_db + shimmer_APQ5 + NHR +
##     HNR + DFA, data = dd2)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10.14344  -1.59523   0.07732   1.56885  10.90455
##
## Coefficients:
##              Value     Std. Error t value
## (Intercept)    40.5604    0.9295    43.6351
## pacient2      -23.9402    0.3600   -66.5015
## pacient3       -6.9004    0.3018   -22.8608
...

...
## pacient41       2.7683    0.3198     8.6576
## pacient42      -7.0520    0.3027   -23.2950
## jitter_pct   -146.6753   57.1988    -2.5643
## jitter_abs   6727.9995 2915.4198     2.3077
## jitter_RAP     95.4533   72.6415     1.3140
## jitter_PPQ5   156.9496   52.6504     2.9810
## shimmer_db      1.2820    0.7933     1.6161
## shimmer_APQ5  -23.7322   10.7693    -2.2037
## NHR            -1.9674    1.7422    -1.1292
## HNR             0.0454    0.0203     2.2338
## DFA            -2.1019    1.0900    -1.9283
##
## Residual standard error: 2.342 on 5824 degrees of freedom
```

De la mateixa forma que s'ha fet amb els dos mètodes anteriors, es construirà una taula per comparar les estimacions dels paràmetres d'aquest model amb els del model de regressió lineal:

Taula 6.3. Comparativa dels coeficients de la regressió lineal i la regressió per mínims quadrats ponderats.

	Regressió Lineal	**Mínims Quadrats Ponderats**
Jitter_pct	-200.72	-146.6753
Jitter_abs	9096.53	6727.9995
Jitter_RAP	131.58	95.4533
Jitter_PPQ5	212.01	156.9496

Shimmer_db	1.91	1.2820
Shimmer_APQ5	-30.95	-23.7322
NHR	-3.29	-1.9674
HNR	0.06	0.0454
DFA	-1.67	-2.1019

Després d'observar les estimacions d'aquest darrer model de regressió robusta, es pot arribar a la mateixa conclusió que s'ha arribat amb els dos models anteriors: les estimacions del model de regressió robusta (en aquest cas de la regressió per mínims quadrats ponderats) són més petites en valor absolut que les del model de regressió lineal com a conseqüència de la menor influència que tenen les observacions atípiques en aquest tipus de regressió.

Finalment, si es comparen mitjançant l'AIC els quatre models obtinguts s'obté:

Taula 6.4. Comparativa dels quatre models mitjançant l'AIC

Model	AIC
Model regressió lineal	28651.25
Model regressió quantil lineal	28408.77
Model regressió robusta T	24854.76
Model regressió mínims quadrats ponderats	28729.12

Es pot comprovar que el model que obté un menor valor d'AIC és el de la regressió robusta T, per tant, seria aquest el més indicat per predir futurs valors de la variable resposta "total_UPDRS".

7. CONCLUSIONS

Per posar fi al treball s'exposaran les conclusions extretes a partir de les aportacions teòriques que s'han introduït i dels resultats obtinguts després de dur a terme l'anàlisi pràctic amb R.

En primer lloc, s'ha arribat a la conclusió que la robustesa és una propietat que cal tenir molt en compte a l'hora d'escollir un estimador estadístic. A partir d'un petit exemple en que s'han calculat la mitjana i la mediana d'una mostra que contenia un valor atípic i s'han comparat els resultats d'ambdós estimacions amb els de la mateixa mostra, excloent aquesta dada que no es comporta de la mateixa forma que la resta, s'ha comprovat que el valor de la mitjana de la mostra que conté el valor atípic difereix molt del de la mostra que no el conté, cosa que no passa amb la mediana. És sabut que aquests dos estimadors calculats són mesures de centralització que estimen el mateix paràmetre poblacional (la mitjana poblacional, la qual es denota per μ). Per tant, es pot concluir que, donada la situació en que apareguin valors atípics en una mostra d'estudi, cal tenir molt en compte la robustesa que tenen els possibles estimadors de la variable d'interès que es precisa estimar. En el cas de les mesures de centralització, tal i com s'acaba d'explicar, la mediana seria una millor opció que la mitjana mostral per estimar la mitjana poblacional. I per altra banda, per posar un altre exemple, en les mesures de dispersió, per estimar la variància poblacional hi ha un estimador alternatiu a la variància mostral que s'anomena desviació absoluta de la mediana (MAD) que presenta una major robustesa en front de les dades atípiques.

En segon lloc, coneguda la gran vulnerabilitat que té la regressió lineal per mínims quadrats ordinaris en front els valors atípics, s'ha traslladat el concepte de robustesa a l'àmbit de la regressió, de forma que s'han proposat diversos mètodes alternatius al dels mínims quadrats ordinaris per tal d'obtenir unes estimacions dels paràmetres de la regressió més robustes i d'aquesta forma, poder utilitzar-les per predir qualsevol variable d'interès sense preocupar-se de les més que possibles distorsions (o biaixos) que s'obtenen a la regressió lineal per mínims quadrats ordinaris en situacions en que els valors atípics estan presents a la mostra d'estudi. Aquests mètodes alternatius al dels mínims quadrats ordinaris han sigut el de la regressió quantil lineal, el de la regressió robusta t i el dels mínims quadrats ponderats. Tots tres mètodes han estat implementats primerament en una mostra d'estudi petita (una variable d'interès i una explicativa, totes dos amb 120 dades generades amb l'R) amb uns quants valors atípics per comprovar de forma visual la diferència entre la recta de regressió lineal per mínims quadrats ordinaris i la recta de regressió del mètode adient. Després d'això s'ha traslladat l'exemple a un problema més pràctic ja que s'ha utilitzat una mostra que perfectament es podria trobar a un estudi estadístic de la vida real on s'utilitzaven al voltant de 10 variables explicatives per predir-ne una dinterès, i d'aquesta forma se n'ha comprovat la potència dels mètodes utilitzant dades reals. L'aplicació dels mètodes de regressió robusta amb la mostra petita de prova ha resultat tot un èxit ja que la diferència entre les rectes de regressió robusta i la de regressió lineal per mínims quadrats ordinaris ha sigut més que evident, sent el mètode de la regressió robusta t el que ha obtingut una recta de regressió que

s'ha vist menys distorsionada pels valors atípics. En quant a l'exemple amb el conjunt de dades més gran, com que s'ha construït un model de regressió amb 9 variables explicatives (i també l'indicador de cada pacient com a bloc, ja que hi havia més d'una observació per pacient), no s'ha pogut comprovar la diferència entre les estimacions dels paràmetres de cada model de forma gràfica tal i com s'ha fet abans amb els models on només hi havia una variable explicativa per motius de dimensionalitat. Per tant, en aquest cas s'ha realitzat la comparativa entre el model de regressió lineal per mínims quadrats ordinaris i els models de regressió robusta estudiats directament amb els valors de les estimacions dels seus paràmetres i s'ha arribat a la conclusió que els models de regressió robusta obtenen unes estimacions dels paràmetres més petites (en valor absolut) que el model de regressió lineal per mínims quadrats ordinaris com a conseqüència de la menor influència que es dóna als valors atípics.

En definitiva, els mètodes de regressió robusta obtindran sempre unes estimacions dels paràmetres més moderades que el mètode de la regressió lineal per mínims quadrats ordinaris, com a conseqüència de donar-li una major importància al conjunt de dades central que no pas als valors atípics situats als extrems de la mostra.

8. BIBLIOGRAFIA

A Tsanas, MA Little, PE McSharry, *Accurate telemonitoring of Parkinson.s disease progression by non-invasive speech tests* [en línia]. LO Ramig, 2009. IEEE Transactions on Biomedical Engineering (to appear). Disponible a: <http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring>.

Carmona, Francesc. *Modelos ineales*. Barcelona: Publicacions i Edicions de la Universitat de Barcelona, 2005. ISBN 9788447528936.

E. Ramalle- Gómara, JM. Andrés De Llano. "Utilización de métodos robustos en la estadística inferencial". Atención Primaria. 32(3):177-82, 2003.

Estefany Melissa Minalla Alava, Mario David Solórzano Carvajal, Msc. Gaudencio Zurita Herrera. *Construcción de Software para Regresión. El caso de Regresión Ridge y Robusta.* Instituto de Ciencias Matemáticas. Campus Gustavo Galindo, Km 30.5 vía Perimetral. Guayaquil-Ecuador.

Ford, Clay. *Statistical Research Consultant* [en línia]. University of Virginia Library, 20 de Setembre del 2015. Disponible a: <https://data.library.virginia.edu/getting-started-with-quantile-regression/>.

H. Zamar, Ruben. Estimación robusta. A: *Estadística española*. 1994, vol. 36, núm. 137, p. 327-387. ISSN 0014-1151.

Kenneth L. Lange, Roderick J. A. Little and Jeremy M. G. Taylor. Robust Statistical Modeling Using the t Distribution. A: *Journal of the American Statistical Association*. 1989, vol. 84, núm. 408, p. 881-896.

Kjytay. "Quantile regression in R". R-bloggers. 31 de Gener de 2019. Disponible a: <https://www.r-bloggers.com/quantile-regression-in-r-2/>.

N. Areshenkoff, Corson. "Robust t-regression". areshenkBlog. 1 de Novembre de 2014. Disponible a: <http://areshenk-research-notes.com/robust-t-regression/>.

9. ANNEXOS

```
########################################
#APARTAT 2.3. ROBUSTESA D'UN ESTIMADOR#
########################################

set.seed(3)

x <- rnorm(n = 25,mean = 50,sd = 2)
stripchart(x,method = 'overplot',pch = 16,group.names = 2,main = "Dades sense valors
atípics",xlim=c(46,53))

mean(x)
median(x)



y <- c(x,70)
stripchart(y,method = 'overplot',pch = 16,group.names = 2,main = "Dades amb valors
atípics",xlim=c(45,70))

mean(y)
median(y)


#################################
#APARTAT 4.1. REGRESSIÓ LINEAL #
#################################

#DADES LINEALS

x <- c(1:10)
y <- 5*x
plot(x,y,pch=20,xlab="",ylab="")
abline(lm(y~x))
title("Exemple Linealitat")

#DADES NO LINEALS

x <- c(-10:10)
y <- x^2
plot(x,y,pch=20,xlab="",ylab="")
abline(lm(y~x))
title("Exemple No Linealitat")

#DADES HETEROSCEDASTICITAT

set.seed(23)

data <- mvrnorm(n = 100, mu = c(5,5), Sigma = matrix(c(1, 0.99,0.99,1),ncol=2))
data <- data.frame(data);data <- data[data$X1 < 5.5,]

data2 <- mvrnorm(n = 50, mu = c(6.25,6.25), Sigma = matrix(c(1, 0.7,0.7,1),ncol=2))
```

```
data2 <- data.frame(data2);data2 <- data2[data2$X1 >= 5.5 & data2$X1 <8,]

data3 <- mvrnorm(n = 50, mu = c(7.5,7.5), Sigma = matrix(c(1, 0.6,0.6,1),ncol=2))
data3 <- data.frame(data3);data3 <- data3[data3$X1 >=8,]

dd <- rbind(data,data2,data3)


names(dd) <- c("X","Y")
plot(Y~X,dd,main="EXEMPLE HETEROSCEDASTICITAT",pch=20,xlab="",ylab="")


#DADES HOMOSCEDASTICITAT

set.seed(23)

data <- mvrnorm(n = 100, mu = c(5,5), Sigma = matrix(c(1, 0.99,0.99,1),ncol=2))
ddd <- data.frame(data)
names(ddd) <- c("X","Y")
plot(Y~X,ddd,main="EXEMPLE HOMOSCEDASTICITAT",pch=20,xlab="",ylab="")

################################################
#APARTAT 4.2. APLICACIO REGRESSIO ROBUSTA##
################################################

library(MASS)
library(quantreg)

set.seed(51)

n1 <- 100
n2 <- 20

#generem 1000 observacions "bones"

data <- mvrnorm(n = n1, mu = c(0,0), Sigma = matrix(c(1, 0.9,0.9,1),ncol=2))

#generem 50 observacions "dolentes" o "outliers"

data <- rbind(data, mvrnorm(n = n2, mu = c(1.5,-1.5), Sigma = matrix(c(0.2,
0,0,0.2),ncol=2)))


data <- data.frame(data)
names(data) <- c("X","Y")
ind <- c(rep(1,n1),rep(2,n2))

plot(Y~X, data, pch=c(20,16)[ind], col= c("black","red")[ind], main = "Influència dels
valors atípics a la regressió")


abline(lm(Y~X,data),lty=2,col="red")
abline(lm(Y~X,data,subset=1:n1),lty=1,col="blue")
legend("topleft", c("Regressió amb dades 'bones'","Regressió amb totes les dades"),
      inset=0.01, lty=c(1,2),col=c("blue","red"),cex=0.7)
```

```
###################################
#APARTAT 5.1. REGRESSIÓ QUANTIL#
###################################

library(MASS)
library(quantreg)

set.seed(51)

n1 <- 100
n2 <- 20

#generem 1000 observacions "bones"

data <- mvrnorm(n = n1, mu = c(0,0), Sigma = matrix(c(1, 0.9,0.9,1),ncol=2))

#generem 50 observacions "dolentes" o "outliers"

data <- rbind(data, mvrnorm(n = n2, mu = c(1.5,-1.5), Sigma = matrix(c(0.2,
0,0,0.2),ncol=2)))


data <- data.frame(data)
names(data) <- c("X","Y")
ind <- c(rep(1,n1),rep(2,n2))

plot(Y~X, data, pch=c(20,16)[ind], col= c("black","red")[ind], main = "Dades
generades")
legend("topleft", c("Dades 'bones'","Dades atípiques   "),
      inset=0.01, pch=c(20,16),col=c("black","red"),cex=0.9)



plot(Y~X, data, pch=c(20,16)[ind], col= c("black","red")[ind], main = "Exemple
Regressió Quantil")

summary(r1 <- rq(Y~X, data=data, tau=0.5))
abline(r1)
abline(lm(Y~X,data),lty=2,col="red")
abline(lm(Y~X,data,subset=1:n1),lty=1,col="blue")
legend("topleft", c("Reg.Lineal OK","Reg.Quantil","Reg.Lineal"),
inset=0.01, lty=c(1,1,2),col=c("blue","black","red"),cex=0.7)



#GRAFIC AMB RECTES DE REGRESSIO PER A DIFERENTS CUANTILS

library(MASS)
library(quantreg)

set.seed(51)

n1 <- 100
```

```
n2 <- 20

#generem 1000 observacions "bones"

data <- mvrnorm(n = n1, mu = c(0,0), Sigma = matrix(c(1, 0.9,0.9,1),ncol=2))

#generem 50 observacions "dolentes" o "outliers"

data <- rbind(data, mvrnorm(n = n2, mu = c(1.5,-1.5), Sigma = matrix(c(0.2,
0,0,0.2),ncol=2)))


data <- data.frame(data)
names(data) <- c("X","Y")
ind <- c(rep(1,n1),rep(2,n2))

plot(Y~X, data, pch=c(20,16)[ind], col= c("black","red")[ind], main = "Exemple Diversos
Quantils")

multirq <- rq(Y~X, data=data, tau = c(0.1,0.3,0.5,0.7,0.9))


colors <- c("#ffe6e6", "#ff9999","#ff0000", "#800000", "#000000")

for (j in 1:ncol(multirq$coefficients)) {
  abline(coef(multirq)[, j], col = colors[j])
}
legend("topleft", c(expression(tau~" = 0.1"),expression(tau~" = 0.3"),expression(tau~"
= 0.5"),expression(tau~" = 0.7"),expression(tau~" = 0.9")),
       inset=0.01, lty=c(1,1,1,1,1),col=c("#ffe6e6", "#ff9999","#ff0000", "#800000",
"#000000"),cex=0.7)


################################
###APARTAT 5.2. REGRESSIÓ T ###
################################

#GRAFIC DISTRIBUCIO NORMAL I T-STUDENT AMB DIFERENTS GRAUS DE
LLIBERTAT

t1 <- curve(dt(x,1), xlim = c(-6,6))
t2 <- curve(dt(x,2), xlim = c(-6,6))
t5 <- curve(dt(x,5), xlim = c(-6,6))


curve(dnorm(x, 0, 1),xlim = c(-6,6),ylab = "",xlab = "")
lines(t1,col=2)
lines(t2,col=3)
lines(t5,col=4)
legend("topleft",c("Normal",expression(nu~" = 1"),expression(nu~" =
3"),expression(nu~" = 5")),inset=0.01,lty=c(1,1,1,1),col =
c("black","red","green","blue"),cex=0.7)
title("Normal vs t-Student")
```

```
#GRAFIC REGRESSIO T AMB DADES OUTLIERS

library(SMIR) #NO VA, FAREM LA FUNCIO A MA (EXTRETA D'INTERNET ->
https://github.com/cran/SMIR/blob/master/R/treg.R)

treg <- function(lm.object, r, verbose){
  if (class(lm.object)!="lm") stop("model must be class ``lm''")
  X <- model.matrix(lm.object)
  y <- lm.object$model$y
  nu <- length(y)
  w <- rep(1,nu)
  d <- 0
  convergence.criteria <- 0.0001
  converged <- FALSE
  fit <- lm.object
  while (!converged) {
    rss <- sum(w*resid(fit)^2)
    phi <- (r + 1) * rss/nu
    sigma <- sqrt(phi/r)
    t <- resid(fit)/sqrt(phi)
    w <- 1/(1+t^2)
    e <- -2*(nu*( lgamma((r+1)/2) -0.5*log(pi) - lgamma(r/2) -
                  0.5*log(phi)) + (r + 1)/2*sum(log(w)))
    if (verbose)cat("-2 log Lmax =",e," with sigma = ",sigma," and scale parameter
",r,"\n")
    converged <- abs(d-e) < convergence.criteria
    d <- e
    fit <- lm.wfit(X,y,w)
    #    cat(r,"\n")
  }
  fit <- c(lm.object,  list(weights = w, disparity=e, tcoef = coef(fit), r=r, sigma = sigma))
  class(fit) <- c("lm","treg")
  fit
}
#
summary.treg <- function(object, ...)
  #  summary.lm
  #function (object, correlation = FALSE, symbolic.cor = FALSE,
  #    ...)
{
  z <- object
  p <- z$rank
  Qr <- object$qr
  #    if (is.null(z$terms) || is.null(Qr))
  #        stop("invalid 'lm' object:  no 'terms' nor 'qr' component")
  n <- NROW(Qr$qr)
  rdf <- n - p
  if (is.na(z$df.residual) || rdf != z$df.residual)
    warning("residual degrees of freedom in object suggest this is not an \"lm\" fit")
  p1 <- 1:p
  r <- z$residuals
  f <- z$fitted
  w <- z$weights
  mss <- if (attr(z$terms, "intercept")) {
    m <- sum(w * f/sum(w))
```

```
  sum(w * (f - m)^2)
 }
 else sum(w * f^2)
 rss <- sum(w * r^2)
 r <- sqrt(w) * r
 #
 resvar <- rss/rdf
 R <- chol2inv(Qr$qr[p1, p1, drop = FALSE])
 se <- sqrt(diag(R) * resvar)
 est <- z$coefficients[Qr$pivot[p1]]
 tval <- est/se
 ans <- z[c("call", "terms")]
 ans$residuals <- r
 ans$coefficients <- cbind(est, se, tval, 2 * pt(abs(tval),
                                    rdf, lower.tail = FALSE))
 dimnames(ans$coefficients) <- list(names(z$coefficients)[Qr$pivot[p1]],
                           c("Estimate", "Std. Error", "t value", "Pr(>|t|)"))
 ans$aliased <- is.na(coef(object))
 ans$sigma <- sqrt(resvar)
 ans$df <- c(p, rdf, NCOL(Qr$qr))
 if (p != attr(z$terms, "intercept")) {
  df.int <- if (attr(z$terms, "intercept"))
    1
  else 0
  ans$r.squared <- mss/(mss + rss)
  ans$adj.r.squared <- 1 - (1 - ans$r.squared) * ((n -
                                      df.int)/rdf)
  ans$fstatistic <- c(value = (mss/(p - df.int))/resvar,
                numdf = p - df.int, dendf = rdf)
 }
 else ans$r.squared <- ans$adj.r.squared <- 0
 ans$cov.unscaled <- R
 dimnames(ans$cov.unscaled) <- dimnames(ans$coefficients)[c(1,
                                         1)]
 #    if (correlation) {
 #        ans$correlation <- (R * resvar)/outer(se, se)
 #        dimnames(ans$correlation) <- dimnames(ans$cov.unscaled)
 #        ans$symbolic.cor <- symbolic.cor
 #    }
 print(ans$call)
 print(ans$coefficients)
 cat("Disparity = ",signif(z$disparity,5),"\n")
 cat("r value   = ", z$r,"\n")
 if (!is.null(z$na.action))
  ans$na.action <- z$na.action
 class(ans) <- c("summary.lm","summary.treg")
 ans
}
```

```
#################################
### APARTAT 5.2. REGRESSIÓ T ###
#################################
```

```
set.seed(51)

n1 <- 100
n2 <- 20

#generem 1000 observacions "bones"

data <- mvrnorm(n = n1, mu = c(0,0), Sigma = matrix(c(1, 0.9,0.9,1),ncol=2))

#generem 50 observacions "dolentes" o "outliers"

data <- rbind(data, mvrnorm(n = n2, mu = c(1.5,-1.5), Sigma = matrix(c(0.2,
0,0,0.2),ncol=2)))


data <- data.frame(data)
names(data) <- c("x","y")
ind <- c(rep(1,n1),rep(2,n2))

plot(y~x, data, pch=c(20,16)[ind], col= c("black","red")[ind], main = "Exemple
Regressió T")

normalModel <- lm(y ~ x,data = data)
tModel <- treg(normalModel, r = 1.1,verbose = F)

abline(tModel$tcoef)
abline(lm(y~x,data),lty=2,col="red")
abline(lm(y~x,data,subset=1:n1),lty=1,col="blue")
legend("topleft", c("Reg.Lineal OK","Regressió T","Reg.Lineal"),
      inset=0.01, lty=c(1,1,2),col=c("blue","black","red"),cex=0.7)




###################################
#APARTAT 5.3. REGRESSIÓ PESOS H#
###################################


set.seed(51)

n1 <- 100
n2 <- 20

#generem 1000 observacions "bones"

data <- mvrnorm(n = n1, mu = c(0,0), Sigma = matrix(c(1, 0.9,0.9,1),ncol=2))

#generem 50 observacions "dolentes" o "outliers"

data <- rbind(data, mvrnorm(n = n2, mu = c(1.5,-1.5), Sigma = matrix(c(0.2,
0,0,0.2),ncol=2)))
```

```
data <- data.frame(data)
names(data) <- c("x","y")
ind <- c(rep(1,n1),rep(2,n2))

plot(y~x, data, pch=c(20,16)[ind], col= c("black","red")[ind], main = "Regressió per
mínims quadrats ponderats")

normalModel <- lm(y ~ x,data = data)
huber <- rlm(y~x,data=data)

abline(huber)
abline(lm(y~x,data),lty=2,col="red")
abline(lm(y~x,data,subset=1:n1),lty=1,col="blue")
legend("topleft", c("Reg.Lineal OK","Mínims Quadrats Ponderats","Reg.Lineal"),
      inset=0.01, lty=c(1,1,2),col=c("blue","black","red"),cex=0.7)



 ###################################
 ### ANALISI DADES PARKINSON #####
 ###################################



library(readr)
dd <- read_csv("~/Desktop/DADES TFG/parkinson.txt")
sapply(dd,class)

dd$`subject#` <- as.factor(dd$`subject#`)

names(dd)
names(dd) <- c("pacient","age","sex","test_time",
            "motor_UPDRS","total_UPDRS","jitter_pct",
            "jitter_abs","jitter_RAP","jitter_PPQ5",
            "jitter_DDP","shimmer","shimmer_db",
            "shimmer_APQ3","shimmer_APQ5","shimmer_APQ11",
            "shimmer_DDA","NHR","HNR","RPDE","DFA","PPE")

#dd$sex <- as.factor(dd$sex)

#sex --> 0 hombre, 1 mujer

#test_time --> intervalo de tiempo desde la fecha de reclutamiento de referencia en
dias.

#variable respuesta --> total_UPDRS

#variables regressoras --> 16 de despues de total_UPDRS --> medidas de voz
biomedicas

#jitter_% - jitter_DDP --> Varias medidas de variación en frecuencia fundamental

#shimmer - shimmer_DDA --> Varias medidas de variación de amplitud.
```

```
#NHR, HNR --> Dos medidas de relación de ruido a componentes tonales en la voz.

#RPDE --> Una medida de complejidad dinámica no lineal.

#DFA --> Señal fractal escala exponente

#PPE --> Una medida no lineal de variación de frecuencia fundamental.

#######QUINES VARIABLES TENEN NULS ? ###########

dd2 <- dd[,c(1,6:22)] #Només les variables a introduir al model

for(i in 1:ncol(dd2)){print(sum(is.na(dd2[,i])))} #No hi ha cap NA, mirare si els nuls
estan en un altre format

sum(table(dd2$pacient)) #Cap nul als pacients

sapply(dd2,summary)

####### NO HI HA NULS #######


#########################################################################
######################## MODEL LINEAL (amb el bloc del pacient) ########
#########################################################################

model1 <- lm(total_UPDRS ~ pacient + jitter_pct + jitter_abs + jitter_RAP +
           jitter_PPQ5 + jitter_DDP + shimmer + shimmer_db +
           shimmer_APQ3 + shimmer_APQ5 + shimmer_APQ11 +
           shimmer_DDA + NHR + HNR + RPDE + DFA + PPE,dd2)
summary(model1)
plot(model1)
library(car)
Anova(model1,type = "III")


##### PROCES DE SELECCIO DE VARIABLES #####

###### BIC ########

model2<- step(model1,direction="both",k = log(nrow(dd)),trace = 0)
terms(model2) #### total_UPDRS ~ pacient + HNR
summary(model2)
plot(model2)


####### AIC ######

model3<- step(model1,direction="both",k = 2,trace = 0)
terms(model3) #### total_UPDRS ~ pacient + jitter_pct +
           #### jitter_abs + jitter_RAP + jitter_PPQ5 +
           #### shimmer_db + shimmer_APQ5 + NHR + HNR + DFA
summary(model3)
plot(model3)
```

```
AIC(model1);AIC(model2);AIC(model3)
BIC(model1);BIC(model2);BIC(model3)
anova(model2,model3)


# Ens quedem en model 3 (AIC)


##MODEL LINEAL

m1_lm <- lm(total_UPDRS ~ pacient + jitter_pct +
          jitter_abs + jitter_RAP + jitter_PPQ5 +
          shimmer_db + shimmer_APQ5 + NHR + HNR + DFA,dd2)

##MODEL DE REGRESSIO ROBUSTA

####REGRESSIO PER QUANTILS (QUANTILE REGRESSION)

library(quantreg)

m2_q <- rq(total_UPDRS ~ pacient + jitter_pct +
            jitter_abs + jitter_RAP + jitter_PPQ5 +
            shimmer_db + shimmer_APQ5 + NHR + HNR + DFA,data = dd2) ####TAU = 0.5
--> MEDIANA, TAU = 0.9 --> PERCENTIL 90%
summary(m2_q)
plot(m2_q$residuals,m2_q$fitted.values)


plot(summary(m2_q), parm="x")

####REGRESSIO REASSIGNANT PESSOS (MIRAR SI ES EL HUBER)

library(MASS)

m3_w <- rlm(total_UPDRS ~ pacient + jitter_pct +
           jitter_abs + jitter_RAP + jitter_PPQ5 +
           shimmer_db + shimmer_APQ5 + NHR + HNR + DFA,data = dd2)
summary(m3_w)

#### T REGRESSION (UTILITZEM LA DISTRIBUCIO T STUDENT)


library(SMIR)
treg <- function(lm.object, r, verbose){
  if (class(lm.object)!="lm") stop("model must be class ``lm''")
  X <- model.matrix(lm.object)
  y <- lm.object$model$total_UPDRS
  nu <- length(y)
  w <- rep(1,nu)
  d <- 0
  convergence.criteria <- 0.0001
  converged <- FALSE
  fit <- lm.object
```

```
  while (!converged) {
    rss <- sum(w*resid(fit)^2)
    phi <- (r + 1) * rss/nu
    sigma <- sqrt(phi/r)
    t <- resid(fit)/sqrt(phi)
    w <- 1/(1+t^2)
    e <- -2*(nu*( lgamma((r+1)/2) -0.5*log(pi) - lgamma(r/2) -
                 0.5*log(phi)) + (r + 1)/2*sum(log(w)))
    if (verbose)cat("-2 log Lmax =",e," with sigma = ",sigma," and scale parameter
",r,"\n")
    converged <- abs(d-e) < convergence.criteria
    d <- e
    fit <- lm.wfit(X,y,w)
    #    cat(r,"\n")
  }
  fit <- c(lm.object,  list(weights = w, disparity=e, tcoef = coef(fit), r=r, sigma = sigma))
  class(fit) <- c("lm","treg")
  fit
}
#
summary.treg <- function(object, ...)
  #  summary.lm
  #function (object, correlation = FALSE, symbolic.cor = FALSE,
  #    ...)
{
  z <- object
  p <- z$rank
  Qr <- object$qr
  #    if (is.null(z$terms) || is.null(Qr))
  #        stop("invalid 'lm' object:  no 'terms' nor 'qr' component")
  n <- NROW(Qr$qr)
  rdf <- n - p
  if (is.na(z$df.residual) || rdf != z$df.residual)
    warning("residual degrees of freedom in object suggest this is not an \"lm\" fit")
  p1 <- 1:p
  r <- z$residuals
  f <- z$fitted
  w <- z$weights
  mss <- if (attr(z$terms, "intercept")) {
    m <- sum(w * f/sum(w))
    sum(w * (f - m)^2)
  }
  else sum(w * f^2)
  rss <- sum(w * r^2)
  r <- sqrt(w) * r
  #
  resvar <- rss/rdf
  R <- chol2inv(Qr$qr[p1, p1, drop = FALSE])
  se <- sqrt(diag(R) * resvar)
  est <- z$coefficients[Qr$pivot[p1]]
  tval <- est/se
  ans <- z[c("call", "terms")]
  ans$residuals <- r
  ans$coefficients <- cbind(est, se, tval, 2 * pt(abs(tval),
                                      rdf, lower.tail = FALSE))
```

```
  dimnames(ans$coefficients) <- list(names(z$coefficients)[Qr$pivot[p1]],
                              c("Estimate", "Std. Error", "t value", "Pr(>|t|)"))
  ans$aliased <- is.na(coef(object))
  ans$sigma <- sqrt(resvar)
  ans$df <- c(p, rdf, NCOL(Qr$qr))
  if (p != attr(z$terms, "intercept")) {
    df.int <- if (attr(z$terms, "intercept"))
      1
    else 0
    ans$r.squared <- mss/(mss + rss)
    ans$adj.r.squared <- 1 - (1 - ans$r.squared) * ((n -
                                          df.int)/rdf)
    ans$fstatistic <- c(value = (mss/(p - df.int))/resvar,
                        numdf = p - df.int, dendf = rdf)
  }
  else ans$r.squared <- ans$adj.r.squared <- 0
  ans$cov.unscaled <- R
  dimnames(ans$cov.unscaled) <- dimnames(ans$coefficients)[c(1,
                                              1)]
  #    if (correlation) {
  #        ans$correlation <- (R * resvar)/outer(se, se)
  #        dimnames(ans$correlation) <- dimnames(ans$cov.unscaled)
  #        ans$symbolic.cor <- symbolic.cor
  #    }
  print(ans$call)
  print(ans$coefficients)
  cat("Disparity = ",signif(z$disparity,5),"\n")
  cat("r value   = ", z$r,"\n")
  if (!is.null(z$na.action))
    ans$na.action <- z$na.action
  class(ans) <- c("summary.lm","summary.treg")
  ans
}

m4_t <- treg(m1_lm, r = 1.1, verbose = F)
m4_t$tcoef

AIC(m1_lm);AIC(m2_q);AIC(m3_w);AIC(m4_t)
```